WILEY PUBLICATIONS IN STATISTICS

The Essentials
# of Educational Statistics

WILEY PUBLICATIONS IN STATISTICS

Walter A. Shewhart *Samuel S. Wilks*
Editors

ALLEN and ELY · International Trade Statistics
BENNETT and FRANKLIN · Statistical Analysis in Chemistry and the Chemical Industry
BLACKWELL and GIRSHICK · Theory of Games and Statistical Decisions
BUSH and MOSTELLER · Stochastic Models for Learning
CLARK · An Introduction to Statistics
COCHRAN · Sampling Techniques
COCHRAN and COX · Experimental Designs
CORNELL · The Essentials of Educational Statistics
CRAMER · The Elements of Probability Theory and Some of Its Applications
DEMING · Some Theory of Sampling
DODGE and ROMIG · Sampling Inspection Tables
DOOB · Stochastic Processes
DWYER · Linear Computations
FELLER · An Introduction to Probability Theory and Its Applications, Volume I
FISHER · Contributions to Mathematical Statistics
FRYER · Elements of Statistics
GOULDEN · Methods of Statistical Analysis, *Second Edition*
HALD · Statistical Tables and Formulas
HALD · Statistical Theory with Engineering Applications
HANSEN, HURWITZ, and MADOW · Sample Survey Methods and Theory, Volumes I and II
HAUSER and LEONARD · Government Statistics for Business Use, *Second Edition*
HOEL · Introduction to Mathematical Statistics, *Second Edition*
KEMPTHORNE · The Design and Analysis of Experiments
MUDGETT · Index Numbers
RAO · Advanced Statistical Methods in Biometric Research
RICE · Control Charts
ROMIG · 50 100 Binomial Tables
SAVAGE · Foundations of Statistics
TIPPETT · The Methods of Statistics, *Fourth Edition*
TIPPETT · Technological Applications of Statistics
WALD · Sequential Analysis
WALD · Statistical Decision Functions
WOLD and JURÉEN · Demand Analysis
YOUDEN · Statistical Methods for Chemists

The Essentials of Educational Statistics

FRANCIS G. CORNELL
Professor of Education,
University of Illinois

New York · John Wiley & Sons, Inc.
London · Chapman & Hall, Limited

Library of Congress Catalog Card Number: 56-6482

PRINTED IN THE UNITED STATES OF AMERICA

Preface

THIS BOOK IS WRITTEN FOR STUDENTS OF EDUCATION AS AN INTRODUCTION to the subject of statistics. It began as notes which I had accumulated in the course of studying practical problems in the operation of schools and in educational research, and as supplements which I had prepared in my classes in educational statistics. In my teaching I have found it necessary to use textbooks which were in some respects unsatisfactory. The best available were usually not addressed exclusively to the student of education. Some I found in one way or another to be too advanced, technically unsound, or shallow in their treatment of theory and the underlying concepts of the subject.

There has been marked improvement in recent years in both introductory and advanced statistics textbooks in general. I have noted this improvement in at least these categories: mathematical statistics texts, applied statistics texts for fields not directly related to education (viz., agriculture), texts on special applications of statistical methods of interest primarily to advanced students in education (viz., survey sampling methods), and statistics texts for students in psychology *and* education. Several such texts are included among references in this book. A selection from among them along with an appropriate file of journals would constitute an excellent reference library for students of educational statistics. Very little in the present volume could not be found in such a collection. Such justification as I have in writing this book is hence not for the novelty of its substance but for having presented it and organized it from the vantage point of problems in education and from the vantage point of the background of the student in education.

It has been common practice for some time in education to cover so-called "descriptive statistics" during the first semester or two of instruction, reserving statistical inference for advanced instruction. As a consequence a majority of students have never got to the heart of the subject. It is an unrealistic way of splitting the subject of educational statistics, it is uneconomical since usually it means that students must learn two or more contradictory systems, and it is pedagogically unjustified on either logical or psychological grounds.

In planning this book I was strongly tempted to launch the subject with a sample problem in the first chapter. I dropped this notion in favor of giving first attention to a review of the necessary elements of arithmetic and algebra, and providing early experience in the manipulation of data. This decision seems warranted in view of the limited mathematical training of most education students, and the apparent advantages for such students of an intuitive presentation with concrete numerical examples. When the student deficient in simple numerical operations is not given remedial instruction, he spends his time on computation and not on learning statistics. The first four chapters are thus quite conventional in approach, covering the usual topics through measures of central tendency and dispersion. From that point on, however, this book is grounded in sample theory.

As an *introduction* to statistics, this book is not a *handbook*, though it contains more than enough material for the usual first two semesters of instruction. The intent has been to include *essentials* both of specific techniques and of underlying ideas. It seems reasonable that a book with this purpose should cover a limited ground thoroughly rather than much ground lightly. The reader will not find some topics which I have been tempted to include, such as survey sampling methods, covariance analysis, and discriminate analysis; nor will he find an extensive treatment of many practical techniques. The publisher's reviewers have agreed with me that such topics are not necessary for my objective, particularly since they are effectively treated elsewhere. It is my hope that a mastery of this book will prepare a student for further study of these and other topics in statistics. Their omission permits the inclusion of an elementary treatment of the underlying theory on basic topics in statistics not usually found in introductory textbooks.

One objective has been to enable the student of education to look at statistics a little more as statisticians themselves are now looking at it. It is no simple matter to make a mathematical subject easy to learn. Some mathematical ideas are not readily converted to simple language. A book of this type is limited in space, and the writer must restrict himself in the number of displays and examples he uses. One hazard he faces is failure to bring the subject matter down to the proper level; another is loss of the elegance and precision of the mathematics. Where I have not successfully overcome these hazards, I trust that supplementary reading and instruction will be employed if individual users find it necessary.

I am greatly indebted to Professor Horace W. Norton for a painstaking reading of the entire manuscript and to Professors Eric F. Gardner and Lincoln E. Moses who reviewed the first draft of this book. Their comments pointed out a number of technical errors and ambiguities which have been eliminated. I am grateful to C. Henry Carlson for assistance in the final technical editing of the manuscript. My thanks are also due my former assistants and many former students whose cooperation in the use of preliminary drafts of this book in my classes at the University of Illinois was of inestimable value.

I am indebted to Professor Sir Ronald A. Fisher, Cambridge, and to Messrs. Oliver and Boyd, Ltd., Edinburgh, for permission to reprint Table III from their book *Statistical Methods for Research Workers;* to Professor George W. Snedecor and the Iowa State College Press for permission to publish Snedecor's table for the F distribution from *Statistical Methods;* to Professor John F. Kenney and the D. Van Nostrand Company for permission to reproduce from *Mathematics of Statistics* the table of Appendix C; to Alexander M. Mood and the McGraw-Hill Book Company for permission to reproduce from *Introduction to the Theory of Statistics* the table of Appendix D; to the Rand Corporation and *The Journal of the American Statistical Association* for permission to reproduce the table of random numbers in Appendix B; to Helen M. Walker and Walter M. Durost and the Bureau of Publications, Teachers College, Columbia University, for permission to adapt materials from *Statistical Tables, Their Structure and Use;* and to W. Edwards Deming, Carl W. Proehl, Harold E. Mitzel, Louis P. Aikman, and William C. Krathwohl for their kindness in permitting the adaptation of their materials.

FRANCIS G. CORNELL

University of Illinois
June, 1955

Contents

CHAPTER 1

The Contribution of Statistics in Education

In common conversations the term *statistics* refers to sets of figures. We have occasion to refer to population statistics in census reports, enrollment statistics in reports of school systems, or test scores and other types of numerical information about individual pupils—all referred to as *statistics*. This common notion that statistics has to do with figures is correct, but in the study of statistics we find that it is considerably more than simply the manipulation of figures or quantitative data. The study of statistics involves learning about an important scientific tool useful in effectively operating programs of education, in developing the science of education, and in understanding better the complex world in which we live.

1.1 THE MEANING OF STATISTICS

There are at least three meanings of statistics: (*a*) *figures* derived from quantitative data, (*b*) the *process* by which numerical data are treated, and (*c*) *systems of analysis* for interpretation of quantitative data. The statistician uses the term statistics in still another sense: descriptive measures derived from samples from an entire population or universe as distinct from such measures derived from all members of the population or universe.

It is usually easy to tell from context which meaning people have in mind when writing or talking about statistics. The chief reason for bringing up the subject of definition here is that, although we must know about techniques, how to compute, and how to use formulas and such, the emphasis in the study of statistics which we are making is on *method*, that is, statistics as a useful tool for thinking and acting.

1.2 STATISTICS AND MATHEMATICS

We are interested in only an introduction to the subject of statistical method in education. Our treatment will be what is commonly called "nonmathematical," but as we should have sensed already there really is no such thing as nonmathematical statistics. It is merely double talk to say that the ideas of statistics can be understood without mathematics, for in fact the background of *statistical method* is the *theory of statistics*, a field of applied mathematics.

If we become interested professionally in making considerable use of the methods of statistics, it will be necessary for us to study in some detail the mathematical backgrounds of the statistics usable in the field of education. It is assumed that we have a sufficient mastery of elementary mathematical skills and concepts to permit us to understand the elements of common statistical techniques and some of the major ideas underlying them.

Many students in education who do not have a mathematical background become frustrated and emotionally blocked in a way which discourages achievement in their study of statistics. It has been demonstrated, however, that considerable understanding of the elements of the application of statistical ideas can be mastered without advanced mathematics. They will probably need to brush up on rusty skills in elementary arithmetic and some algebra and to take an inventory of their ability in this area. No student competent to do advanced undergraduate or graduate work in education is incapable of an acceptable minimum of proficiency in algebra and computational techniques necessary for successful study of the elements of statistical method.

1.3 WHY STUDY STATISTICS IN EDUCATION?

There are three major reasons why students who are now working in education or who plan to work in education should study statistics. First, a knowledge of statistics is essential nowadays for occupational competency. Technological developments and developments in the science of education and in the application of statistical methods to school problems have made the teacher's, the supervisor's, and the administrator's jobs much more complex than they were half a century ago. They handle more funds. Administration and management operate in many ways requiring an elementary working knowledge of statistical methods. The teacher, the guidance counselor, the curriculum specialist —each is lost without at least the simplest tools of quantitative analysis which will enable him to make use of the vast array of mental and educational tests and other devices of evaluation now available.

The other two major reasons for the study of statistics are matters of *literacy*. A generation or two ago a teacher was equipped for his job if he had literacy of about an eighth-grade level. He knew the subjects of reading, writing, and reckoning as they were taught in those days, and most teachers in the rural "common schools" dealt with youngsters on that level or lower. Indeed, it was not long ago that the standards of literacy for the profession of education were only slightly higher than those now established by our armed forces, which is a literacy level of the fourth grade. But here we are not talking about *verbal* literacy but about *statistical* literacy. Teachers need to be able to read and to understand and to interpret literature from the many disciplines which are drawn upon in the field of education and which are increasingly making use of statistical methods. This is a matter of *professional* literacy.

There is another reason for educators to possess statistical literacy. Our culture has become very largely a statistical culture. Part of the general education of any teacher, therefore, should be an understanding of the systems of thought and the basic concepts of statistical method so that he may be better equipped to understand the world in which he lives. He thus may become better equipped to interpret our culture to young people, as they prepare themselves for citizenship, through a better understanding of ways of handling quantitative or statistical ideas.[1]

For a student who plans to do research in education the importance of statistical method as a research tool is no doubt already clear. Although statistical technique should not be expected to be a panacea with which to work magic in the study of educational problems, its use is becoming more and more important.

1.4 STATISTICS AND SYSTEMATIC PLANNING

A knowledge of the techniques of statistics in *advance* of planning a research study can simplify the task of assimilating large masses of data. In reaching conclusions, the researcher can err by a lack of knowledge of the phenomena in the field of application in which he is working, and he can err as well by an improper use and lack of understanding of the statistical manipulations which he uses.

R. A. Fisher, whose contribution has been great in the development of methods of efficiently handling masses of data, has said: "No human mind is capable of grasping in the entirety the meaning of any considerable

[1] See references 1, 2, and 5 at the end of this chapter.

quantity of numerical data".[1] In a sense, the major contribution of the study of statistics is to develop an understanding of techniques by which the whole of the relevant information may be extracted from data, not only by means of analysis after data have been collected but also by designing and planning the program of measurement and data collection so that a statistical analysis can lead to efficient and dependable interpretation.

1.5 STATISTICS AND EDUCATION

In educational research and in operating educational programs we deal with a variety of problems. Education as a social science is young; yet the development of twentieth century education supports the observation of Adolphe Quetelet, who pioneered the application of nineteenth century probability theory to the social sciences: "The more advanced the sciences have become, the more they have tended to enter the domain of mathematics, which is a sort of center towards which they converge. We can judge the perfection to which a science has come by the facility, more or less great, with which it may be approached by calculation."[2]

As a tool for solving operating problems in schools, statistical method is indispensable. Can you describe the types of applications of statistical method involved in the following?

1. Controlling the maintenance and operation costs for school buildings.

2. Establishing statistical reporting and record systems for controlling school attendance.

3. Analyzing and interpreting school promotions and school progress statistics.

4. Determining costs of different kinds of educational expenditure and making cost analyses.

5. Evaluating school marking systems.

6. Analyzing and interpreting text results for pupil personnel and guidance work.

7. Determining whether or not homework assignments are reasonable.

8. Studying the utilization of library facilities.

9. Answering the question, "How well do we pay our teachers"?

10. Studying class size.

11. Determining whether or not there is adequate light at each pupil's study station.

[1] R. A. Fisher, *Statistical Methods for Research Workers*, New York, Hafner Publishing Co., p. 6.

[2] Quoted from Helen M. Walker, *Elementary Statistical Methods*, New York, Henry Holt and Co., 1943.

1.6 EVERYDAY STATISTICS

On the question of the importance of general statistical literacy, we may examine any issue of any widely read journal, weekly, or monthly, or any newspaper of national coverage to see how many columns, pages, or articles are concerned with not just figures or statistics but statistical method as well. For instance, on the day that this was written, the *New York Times* had five articles on the first page alone treating statistical ideas or methods which could easily be misunderstood by persons not acquainted with the study of statistics. One article had to do with the property tax rate; one reported a survey in 42 colleges and universities on American and world geography; another reported *estimated* casualties in a disaster; another *predicted* a record uranium output in a Canadian field; and still another, dealing with United States economic and financial aid to a foreign government, discusses production of certain important commodities in international trade. Numerous other articles appeared in the same issue of this newspaper, not to mention the financial statistics in the stock market and business sections. Even the sports section, with the "box scores" for major and minor league baseball, gave the standing of the various leagues and reported batting averages.

Do we know how tax rates are computed? Do we know how property is assessed? Do we know what types of errors can come about in the application of a school tax rate in the school district in which we may now be interested? These things are important to educators in a very real sense because local property tax rates yield revenues for schools.

The writer of the *New York Times* geography article said that United States college students "flunk." How do we know how many items and which items on a test students in American colleges should be able to answer? Are there problems of generalizing from only 42 colleges to the total of over 1,800 in the United States? How do we know how many casualties there were? During World War II, according to a postwar bombing survey, including the interrogation of "eye witnesses," bomber crews themselves err greatly in producing estimates of the effect of bombing on the ground and enemy aircraft shot down. When we estimate and predict, what kinds of problems do we have? Such an analysis of news reports of domestic or international events can be made any day in the year, thus illustrating how completely our lives, not as teachers, not as wage earners, but merely as members of a world society, are influenced by complex phenomena which may be understood only with statistical reasoning.

1.7 HISTORICAL BACKGROUND

Historically, statistics is an infant in the family of sciences. The rudiments of mathematical concepts upon which it is built date back to man's earliest attempts to systematize his problems of measuring and counting properties of populations or groups of items, individuals, or observations of phenomena. It was not until the beginning of the eighteenth century, however, that the Swiss mathematician, Jacques Bernoulli, published his treatise on the theory of probability, a subject which had attracted attention chiefly because of the interest in problems of chance in gambling. This work started a development which culminated in the publication in 1812 of Laplace's probability theory, which served as the foundation for mathematical statistics as it developed in the twentieth century.

The earliest statistics in the social sciences were descriptive and enumerative, providing the types of information now common to us in the modern census. Governments needed information on the basis of which to plan taxes and run affairs of *state*. As a matter of fact, the term *statistics* was derived from this development.

In connection with "state" statistics, Adolphe Quetelet in the nineteenth century gave serious attention to the application of the theory of probability to social statistics. By the turn of the century the theory of correlation had been developed. Since that time, developments in statistical theory and its applications to educational and many other problems have accumulated with an accelerating tempo. Studying statistics today is taking up a subject still in the process of development.

Sophisticated applications of statistical methods of education are a product of the current century. Although interest in problems of educational measurement developed in the nineteenth century, educational measurement is a very new field. Sir Francis Galton in the last century suggested the applications of the "normal curve" to the assigning of grades and marks, and the testing of spelling was systematized by J. M. Rice in 1897. The earliest publication of statistical methods applied to educational problems was Thorndike's book, which appeared in 1904.[1] The development of psychological and educational measurement was marked by the Army Alpha Intelligence Test developed in World War I. A series of applications of measurement in education followed at an increasingly rapid rate. There were extensive uses of educational and

[1] E. L. Thorndike, *An Introduction to the Theory of Mental and Social Measurements*, New York, Teachers College, Columbia University, 1904.

psychological testing by the armed forces in World War II for purposes of selection, classification, and training of military personnel.

There are two other ways in which statistics in the last few decades has become of great significance in education. The first began with the survey movement. It was a development of systematic evaluation of an American educational enterprise which had changed in size and complexity from the little red schoolhouse of the last century. With this desire for the systematic study of school systems, it became necessary to improve methods of collecting and analyzing data and measuring many diverse aspects of school systems—not only achievement, but also such matters as the adequacy of school buildings and methods of operation of school boards.

The most recent impetus to the use of statistical methods in education is the development of educational research. Not many years ago educational research was almost exclusively either a philosophical activity, on the one hand, or a straightforward fact-getting and describing activity on the other. With the advent of improved measuring instruments came an interest in experimental methods in education. This has increasingly required more systematic and precise applications of the mathematics of statistics.

As this book was being written, current periodicals contained articles commemorating the twenty-fifth year since the publication of R. A. Fisher's *Statistical Method for Research Workers*. Fisher's book marked the beginning of a series of developments in statistical method. Curiously enough, it is only some twenty-five years since the first so-called scientific study of school finance was made, to cite one field in which statistics has made an important contribution in recent years. This study led to a sequence of studies which, largely because of the power of statistics as a methodological tool, are producing a genuine "theory" of school finance. Other landmarks of advances in education have been achievements in the teaching of reading and arithmetic, all of which progressed rapidly as applications of statistical methods came to be used by students of education.

We are still on the threshold of vast unexplored areas requiring research in education. New concepts, new understandings, new philosophies, new theories of education have called for greater precision, greater insight, sharper technical skill on the part of the empirical investigator in the field. Only with larger numbers of statistically literate teachers, administrators, teacher-trainers, educational researchers in universities and in school systems, are we best able to find means of attaining the very complex goals that are dictated for education by our democratic values, by our modern philosophies, and by our enlightened psychologies.

1.8 LOGIC, COMMON SENSE, AND STATISTICS

The successful application of statistical methods to educational problems depends upon expertness in the subject area of investigation as well as upon knowledge of statistical methods. Nevertheless in undertaking statistical studies, in interpreting and applying statistics, in reading reports of others, and in making decisions based on quantitative data, our effectiveness depends primarily upon statistical understanding and meaning of statistical concepts. This understanding requires considerably more intelligence than the simple numerical manipulations which the unskilled might consider statistics to be. Unintelligent use of the statistical method can lead to very absurd conclusions.

Sometimes quantitative facts are purposely "manipulated" by unscrupulous persons bent on deliberately leading to conclusions in defense of only one side of an argument. Yet, as often, these errors may be the errors of the person on the receiving end of the communication channel—the person responsible as an administrator or as a teacher for interpreting facts and, on the basis of them, taking necessary action. Or it may be the researcher who is off guard for lack of experience with statistical methods or for sheer lack of good common sense and logic. A person who has had a reasonable background in the study of statistics is less inclined to fall into these statistical booby traps.

However, these pitfalls in statistics are often violations of very simple axioms of thought. It is thus with considerable good sense that a recent book has been published on the subject, *How to Lie with Statistics*.[1] Unfortunately the statistical and logical slips which actually occur in daily life are more subtle than the more obvious and humorous examples which dramatize the principles violated. One such example is that of the observer who noted that persons who imbibed sufficient quantities of whiskey and water became intoxicated as did those who consumed sufficient quantities of gin and water. Applying the oversimplified canon of John Stuart Mill, "The Method of Agreement," the unwary observer concluded that since the common element in each circumstance was water, it was the water which produced the intoxication. Only an extreme naïveté concerning biochemistry would prevent one from seeing the "hidden cause" in this illustration. Similarly, only a few adults would be so ignorant of basic verities of life as to conclude that "men have more children than women" from statistics showing that there are 1·8 children in families of Princeton graduates, but only 1·4 children in families of Smith graduates.

[1] See reference 3.

EXERCISES

1. What is the difference between the meaning of statistics as the educational savant might use the term and its meaning for (*a*) the layman and (*b*) the mathematical statistician?

2. List and discuss briefly why students of education should study statistics.

3. What are the advantages of statistical method in educational research?

4. What do you think should be the objectives of a course in statistical methods in education?

5. Can you cite one or more studies in which statistical method contributed to our knowledge of:

(*a*) The use of moving pictures and radio in teaching.
(*b*) The relationship of school to community.
(*c*) Population trends and school enrollment.
(*d*) The influence of family environment on pupil development.
(*e*) Methods of improving reading.
(*f*) Determining readability of written materials.
(*g*) Vocabulary difficulty levels.
(*h*) Diagnostic and remedial methods for individuals having speech or hearing defects.
(*i*) The measurement of public opinion on school issues.
(*j*) Teacher supply and demand.
(*k*) Predicting teaching efficiency.
(*l*) Social structure in the classroom.
(*m*) Allocation of state revenues to local school districts.
(*n*) The implications of school district reorganization.
(*o*) The evaluation of two or more methods of teaching.
(*p*) Selecting materials for the curriculum.
(*q*) The amount and kinds of higher education needed in the United States (or some part of it).
(*r*) Mental hygiene in the school.
(*s*) The measurement of (1) personality, (2) pupil interests and attitudes, (3) motivation, (4) transfer of training, (5) intelligence.
(*t*) Determining desirable school plant characteristics (light, temperature, color, and other aspects of physical environment.)
(*u*) Predicting success in college and determining suitable careers for student guidance.
(*v*) Teaching "critical thinking" or "functional knowledge."
(*w*) Mental development in infancy.
(*x*) Intellectual and cultural factors in aging.
(*y*) Dietary and nutritional influences on mental development.
(*z*) The relation between physical fitness and physique and morphological variation.

6. What are some other areas of "basic knowledge" important in education in which statistical method has made a valuable contribution.

7. Analyze current issues of a national news weekly or a large daily newspaper for frequency of statistical concepts and logical fallacies in their use.

REFERENCES

1. Deming, William E., and Douglas E. Scates, "Education in Statistics for Participation in Current Affairs," *The School Review*, 56: 262–269, May 1948.
2. Deming, William E., and Douglas E. Scates, "The Need for Statistical Education in High School and College," *The Educational Record*, 29: 72–80, January 1948.

3. Huff, Darrell, *How to Lie with Statistics*, New York, W. W. Norton and Co., 1954.
4. Walker, Helen M., "Allergic to Statistics," *NEA Journal*, 43: 419–20, October 1954.
5. Walker, Helen M., "Statistical Literacy in the Social Sciences," *The American Statistician* 5: 6–12, February 1951.
6. Walker, Helen M., "Statistical Understandings Every Teacher Needs," *NEA Journal*, 43: 21–2, January 1954.

CHAPTER 2

Counting and Measuring in Education

It is the purpose of this chapter to distinguish between types of data with which educational people commonly deal, and to point out some of the elementary rules for "processing" such data.

It is important that we develop a sense of the nature of the data with which we are to work. A common failing of students in education is choosing a *method* not suitable for the type of data to be treated.

2.1 COUNTING AND ENUMERATION

Statistical operations designed primarily for description frequently require only measures which are counts of discrete things or objects. A common source of educational data is the basic records of local schools, state departments of education, or the United States Office of Education. These basic records are often abstracted and summarized so as to give *frequencies* or numbers, that is, enumerations of various items of interest to school people.

The commonest items are *number* of pupils or *number* of teachers, a person being the unit of counting. All school districts, all state departments, and the Federal Office of Education, for instance, report the numbers of pupils and the numbers of teachers by grade and often by some geographical unit such as county or state.

Of course there are many ways that we may define *a pupil* or *a teacher* for purposes of such counting. One method of counting pupils is *enrollment*, defined as the total number of names, in the schools or classes, which have been listed at any time during the year. A common method is *average daily attendance*, usually defined as the number of pupil-days attendance divided by the number of days schools are in session. Thus, in a given class, if on one day there were 26 pupils in attendance, on a second 27, and on a third 25, the average daily attendance for the three days would be the result of dividing (26 plus 27 plus 25) by 3, which is equal to 26. Similarly, in the counting of teachers, definitions might vary. It would be important in designing a record system or in planning a survey to decide whether or not to count part-time teachers and teachers who have not been employed for the entire year or for the period of the survey.

Nevertheless, such data, despite difficulties of definition, have one thing in common—theoretically they are *exact counts*. A school district employs an integral number of teachers; their names exist on the payroll; and they can be identified. At a given time, a third-grade teacher has an integral number of boys and girls assigned to her class or in actual attendance. Assuming accurate counting, the annual report of a local board of education will show the exact number of dollars and cents expended during the year for such purposes as transportation, salaries, operation, maintenance, debt service, and capital outlay. Therefore, many statistical operations are intended to yield descriptive information on *how many*.

Enumeration data or information resulting from the process of counting is important not only for the descriptive statistics of records and reports, but also for research and analytic studies in education. In both descriptive statistical work and in analytical research studies, enumeration is usually made according to some meaningful system of classification. For instance, in a study of the processes of the introduction of innovations in schools, various persons were classified in terms of "role" which they had been exercising both in the past and in their current efforts. Seven categories are shown in Table 2.1 for the classification of 249 observations

TABLE 2.1

CLASSIFICATION OF 249 OBSERVATIONS OF SUPERINTENDENTS OF SCHOOLS REGARDING SPECIFIC INNOVATIONS ACCORDING TO ROLE OF SUPERINTENDENT*

Role of Superintendent	Number of Observations
Leadership	12
Support	25
Followership	34
Neutrality	125
Ignorance	48
Divided interest	3
Opposition	2
Total	249

* From Paul R. Mort and Francis G. Cornell, *American Schools in Transition*, New York, Teachers College, Columbia University, 1941, p. 203.

of the roles of superintendents of schools in specific changes needed in schools.

In a sample survey following up trainees who had attended war production training during World War II, it was desired to compare the employment status of trainees from various referral sources prior to training. Table 2.2 shows the results of the survey in which trainees in the sample were classified both by their status prior to receiving training and by their employment status (after training) at the time of the survey.

TABLE 2.2

POST-TRAINING EMPLOYMENT STATUS OF TRAINEES, BY STATUS PRIOR TO TRAINING*

Post-Training Employment Status	Status Prior to Training				
	Total	Unemployed	WPA	NYA	Other Employment
Employed	14,132	5,345	3,279	1,068	4,440
By industry	11,350	4,177	2,880	743	3,550
By armed forces	2,637	1,156	297	307	877
By WPA or NYA	145	12	102	18	13
Not Employed	901	320	224	184	173
Unemployed, seeking work	566	184	166	113	103
Not employed, not seeking work	335	136	58	71	70
Total	15,033	5,665	3,503	1,252	4,613

* From U.S. Office of Education, *Preemployment Trainees and War Production*, Vocational Division, Bulletin 224.

Systems of classification of this type are sometimes called *qualitative*. In such a classification would be students grouped according to sex, or occupation of parent, or color of eyes, or some other physical attribute. Other examples are:

1. Classification of school districts by type of district organization, for example, with or without high schools.

2. The "yeas" and "nays" in the vote on a bond issue.

3. The number of pupils responding to each alternative on a multiple-choice test item.

4. The number of heads and the number of tails in several tosses of a coin.

5. The frequency of occurrence of different combinations of cards, drawn five at a time from a deck.

2.2 MEASUREMENT ON A CONTINUUM

Another very frequent type of data of concern to the educator answers the question, "*How much?*" Thus a student record in most high schools reports date of birth, how much time since birth. At any time *ages* of pupils may be computed. At various times, commonly once a year, physical measurements are recorded. By means of tests we have *scores*, which are measures of *how much* ability, *how much* aptitude, or *how much* achievement boys and girls *have* at the time of testing. These are measures on a *continuum*, even though the actual measures are recorded in intervals. For instance, heights are measured to the *nearest* inch, or weight is measured to the *nearest* pound, but actually there can be an infinite number of gradations of height between the shortest and the tallest in a classroom, or of weight between the lightest and the heaviest.

Scores on tests are frequently the results of counting items. A score of 33 on a test may mean 33 items answered correctly. The underlying variable, intelligence, or achievement of some type, may nevertheless be considered as continuous.

In the physical sciences, in the biological sciences, and in the social sciences measurements of *effects* of properties of objects or persons are commonly used to infer quantitative descriptions of these properties. Changes in temperature are measured by a thermometer which shows the effect of heat upon the expansion or concentration of a liquid or solid. Although it is not within our province here to review theory of measurement, it should be noted that such methods of measuring by indirectly observing and recording effects, symptoms, or correlatives of a characteristic or property frequently require careful study of the assumptions made and the nature of the units of measurement which result.[1]

Frequently it is useful to classify subjects into groups or classes on the basis of continuous measures. This is done, for example, when a qualifying standard or "cut-off" point is established for selection, using test scores or other measures, for such purposes as Selective Service and entrance into educational institutions. In Table 2.3, 931 first-grade pupils are classified into five categories of reading readiness as a result of a reading readiness test and an intelligence test. In this particular instance the individual scores of each of the 931 pupils were available to each teacher for purposes of studying each individual pupil. The basic data

[1] Such questions are usually covered in the study of educational and psychological measurement. Statistical methods covered later in this course are very useful in scaling and measuring problems. Some of the theoretical problems involved in using number systems in measurement are treated in reference 5.

TABLE 2.3

CLASSIFICATION OF 931 FIRST-GRADE PUPILS IN AN EASTERN CITY INTO FIVE READING READINESS GROUPS ON THE BASIS OF A READING READINESS TEST AND AN INTELLIGENCE TEST

Readiness Category	Number of Pupils	Percent
I. Strong capable students	215	23.1
II. Ready to do first-grade work	160	17.2
III. Roughly 50-50 chance of success	232	24.9
IV. Little chance of success in formal work	276	29.6
V. Special treatment	48	5.2
Total	931	100.0

TABLE 2.4

DISTRIBUTION OF MENTAL AGES OF SPRINGFIELD SIXTH-GRADE CHILDREN AS OF APRIL 13, 1948*

Mental Age, Years and Months	Number of Children
15–6 to 15–11	1
15–0 to 15–5	2
14–6 to 14–11	4
14–0 to 14–5	7
13–6 to 13–11	11
13–0 to 13–5	30
12–6 to 12–11	9
12–0 to 12–5	34
11–6 to 11–11	18
11–0 to 11–5	33
10–6 to 10–11	23
10–0 to 10–5	19
9–6 to 9–11	17
9–0 to 9–5	15
8–6 to 8–11	4
8–0 to 8–5	2
7–6 to 7–11	2
Total	231

* From Illini Survey Associates, *A Look at Springfield Schools*, Table 15, p. 109.

from which the table was made, therefore, answered a question of *how much* on two tests—one, reading readiness, and the other, mental ability. However, when the pupils were classified into the five categories on the basis of the test scores, the object was to determine *how many* there were in each category. As a matter of fact, in the analysis of which Table 2.3 is a part, a similar table was prepared for each school and each class in the city so that a count of pupils could be made according to the five readiness categories.

It is often desired to classify individuals on some measure into a convenient number of groups. This is particularly true when dealing with large numbers of subjects. Table 2.4 shows the distribution of mental ages of 231 sixth-grade children tested in a school survey. The advantages of examining this table over examining a tabulation sheet with 231 mental age scores is obvious.

A system of grouping according to half-year intervals was chosen as a matter of convenience. The purpose of the table was to show that the ability levels of sixth-grade classes range over a span of eight years, a span of ability as great as the entire first eight years of school, showing the need for a type of curriculum and for instructional methods which are adapted to the realities of individual differences.

2.3 OTHER DISTINCTIONS IN THE USES OF NUMBERS

The foregoing discussion has made the distinction between enumeration and measurement on a continuum. The discussion may seem unduly theoretical and academic, but its objective is a very practical one. Later it will become clear that vastly different statistical treatment may be required in working with different types of data.

There is another scheme of classification of systems of enumeration with which the educational statistician should be familiar.[1] Series of numerals assigned to objects or events may be called scales. Four types of scales, each with its own characteristic and each requiring its own type of statistical treatment, are:[2]

1. *Nominal.* Assignment of numbers as labels for purposes of classification or identification, for example, "numbering" football players, the use of the Dewey decimal system in classifying books in the library.

2. *Ordinal.* Assignment of numbers to indicate rank or order, for example, rank of the forty-eight states in ability to support education,

[1] S. S. Stevens (Ed.), reference 8.

[2] *Ibid.*

rank of students in a high-school graduating class, order of successive samples of objects in an experiment.

3. *Interval.* Assignment of numbers such that intervals or differences are equal, for example, degrees of temperature (based upon equal volumes of expansion); dates on a calendar; grade scores, age scores, and standard scores, insofar as they successfully "equalize" units.

4. *Ratio.* Assignment of numbers such that equal ratios among them represent equal ratios of some attribute, for example, all forms of counting or enumeration as discussed earlier in this section, such as ages, enrollments of classes; and such *derived* ratio measures as school children per square mile; rates of learning, forgetting, spending school money, etc.

2.4 ARITHMETICAL DEFINITIONS OF UNITS OF MEASUREMENT

As has been suggested above, an important step for any investigator using statistical methods is a definition of his units of measurements, whether he is counting or measuring on a continuous scale. How the investigator defines his units is a question in great part of the subject matter with which he is dealing. It is not primarily a statistical question to determine what it is the investigator wishes to measure, even though statistics may prove very helpful in determining what is measured and how well.

In the study reported in Table 2.2 a distinction was made between persons who were *not employed* and those who were *unemployed.* This decision was made in advance of the survey. During the interviewing of trainees, when they were *not employed*, they were asked whether or not they were *interested in being employed* to the extent of seeking work. Only when it was ascertained that they were seeking work were they classified as unemployed. We note in Table 2.2 that the *not employed* category includes both the unemployed and those who are apparently persons who would not be classed as members of the labor force, those people neither having a job nor seeking employment. This was in accord with the definition of *labor force* in use at the time.

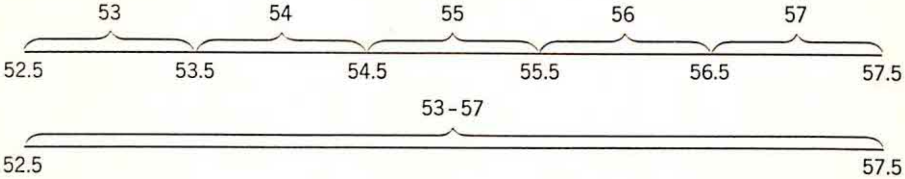
In the study reported in Table 2.3 the question of what measures to use depended upon research in the area of reading and reading readiness. It was not primarily a statistical question. Measures chosen, an intelligence test and a reading readiness test, were those which had been found by students of the teaching of reading in the first grade to be useful predictors of success in first-grade reading.

The number 53 may be used as a measure of the number of students in a class, the number of dollars spent per pupil for a given number of instructional units, the number of towns in a county, the age of the instructor in years, the height of a student in inches, or a score on a test. Regardless of the theory underlying the use of a number, such as 53, there are rules of definition of number which have become more or less standard in statistical work.

Some of the above examples are measures derived from exact counting; others, such as age and height, are "rounded" measures on a continuous scale. For the rounded measurements the scores reported are "approximations." The student is not *exactly* 53 inches in height. In practice, 53 would mean he was nearer to 53 than he was to 52 or 54. The rule usually followed is to record the *nearest* inch or other *nearest* unit.

Following this rule, the student will rarely have difficulty in his computation. The possible exception is the one dealing with age. Customarily in the Western hemisphere, the answer to "How old are you?" is given as "age last birthday." An age of 53 reported in this way means something between the fifty-third birthday and the fifty-fourth birthday. This method of reporting age is avoided in school work by computing from *date of birth* the age to the nearest year or the nearest month. This conforms to the usual convention which we have described.

A series of measures, 53, 54, 55, 56, 57, according to our convention, may be defined graphically by the following diagram:



We see that 53, defined as the "nearest inch, nearest pound, etc.," is considered to be between 52.5 and 53.5; similarly, 54 is between 53.5 and 54.5, and so on, each score representing values on our scale from a point halfway between it and the one below to a point halfway between it and the score above. We may similarly treat scores which are discrete, such as scores on a test. Even though a student has either a score of exactly 53 or a score of exactly some other value which is a single point on the scale, we nevertheless sometimes find it convenient to define 53 as some value between 52·5 and 53·5, and a score of 54 as lying between 53·5 and 54·5.

In order to group scores or to group measures of a number of individuals, such that those with measures of 53 up to and including 57 would be in one category, we would define the limits of the new group of scores as shown in the second line of the above diagram, namely, 52·5 and 57·5.

It is of importance to note, for reasons to be brought out later, that if we were to have several scores or measurements of 53, *to the nearest number*, all of which were proper values lying between 52.5 and 53.5, lacking further information, our best estimate of the "average" value to be assigned these several measures would be 53. This is the *midvalue* of the interval. Similarly, other scores—54, 55, and so on—would be taken to typify a large number of measurements assigned to their respective scale intervals.

In a tabulation in which it was found, for instance, that 25 scores lay in the interval 53–57, the value 55 might be assigned to all the 25 scores, 55 being the *midvalue* of the entire range from 52.5 to 57.5.

We may also note that a method of determining this midvalue or *class mark*, to be used to represent all the measures which have been grouped into an interval, is the lower limit of the interval plus half the size of the interval. In the interval 53–57, for instance, the lower limit of the interval is 52.5. The size of the interval is 57.5 minus 52.5 equals 5.0. One-half of 5.0 is 2.5. Therefore, the class mark is 52.5 plus 2.5 equals 55.0.

2.5 ROUNDING OF NUMBERS

We have suggested above that many measurements are, of practical necessity, *approximations*. The length of a classroom, the width of a table, the height of a student, or the weight of a student must always be reported as an approximation. If the student of education is fortunate enough to be responsible for the original measurements which he is to use, he can measure with the accuracy and precision required for his work. If, on the other hand, he must depend upon the measurements of others, he must take into account the accuracy and precision with which the original measurements were made.

It is good practice in reporting measurements to follow certain rules so as to indicate the precision involved. For instance, if the height of a pupil is reported as 53.46 inches, the simple fact that figures appear in the first and second decimal place should mean that the pupil was measured to the nearest hundredth of an inch. In other words, the individual for whom 53.46 was reported had a height of at least 53.455 inches but less than 53.465 inches. However, the wary investigator would be suspicious of measurements of heights of pupils reported to this degree of accuracy, knowing that differences in posture, time of day, etc., would result in measurement errors much larger than one-hundredth of an inch. In other words, good reporting of measurements involves not only following the convention to which we have referred, but also exercising good judgment in not wasting time with more figures than can be justified for the material and the measuring instruments at hand. Similarly, figures derived from measurements, such as averages and percentages, should not be reported in terms not justified by the original data. For most uses to which the student of education would put a percentage, hundredths of a percent will be more than adequate. In a study of percentage of overageness in the third grade, 23.56 probably would be just as meaningful if reported 23.6. The mean heights of a group of students need not be reported 68.328 inches. Probably 68.3 inches would be enough.

In the two examples just given, the rule for rounding has been applied. It is the convention we have stated in Section 2.4, that is, rounding to the nearest number. In reporting the mean height in inches, we would report 68 if we chose to report to the nearest inch.

If we reported 68.00, the use of the two ciphers is a reporting device to show precision, and, according to the convention, would mean that the observed mean lay between 67.995 and 68.005. In the case of whole numbers, ciphers are used to indicate rounding. For instance, to the nearest dollar, Schenectady, New York, spent $131,114 for adult education in 1946-1947. If this were to be rounded to the nearest thousand dollars, the amount $131,000 would be reported.

Sometimes if a measure should come out by accident exactly an even number of thousands or millions, a period is placed after the last cipher to indicate it is exact and not an approximation. If the expenditure in a school district was $2,800,000.00, the location of the period, and particularly the two ciphers for cents, would show us that it was *exactly* that amount. If it was a rounded figure, it would be reported $2,800,000 without the period at the end of the row of figures.

The student should have no difficulty in applying the rules of rounding to the nearest chosen unit, whether thousands, tenths, thousandths, integers, or what not, if he examines the procedures in the following examples:

Original Measure	Rounded Number
526,786	527,000
14.836	14.8
0.00568	0.006
69.98	70

There is one special case that might prove confusing. Sometimes the first digit to the right of those to be retained is exactly 5, or 5 followed by zeros. For example, suppose percentages were reported in hundredths, and it was desired to round them to tenths, and one of the percentages was 43.65. From our definition of number, 43.65 is exactly on the dividing line between the range that we label 43.6 and the range that we label 43.7. Of course, when the original measurement was taken, it might have been anything between 43.645 and 43.655; but if the original data are not available, we do not know whether it was just a little less than 0.65 or just a little more than 0.65. In such a case, one choice is about as good as another.

It is usual for untrained persons to follow the rule of simply dropping the 5. The disadvantage is that this rule causes under-reporting as a result of lowering every measure ending in five. Presumably in about half the cases like that above, the true values would be nearer 6, and half would be nearer 7. Therefore, a better rule is one which would result in rounding half one way and half the other.

The device which we will use is to *round to the nearest even number*. Following this rule, we would arbitrarily round 43.65 to 43.6. On the other hand, if we were to round 28.35 by this rule, it would become 28.4. This time we are rounding *up*. Since we would expect about half to have odd numbers preceding the 5 and half even numbers preceding the 5, this rule will lead us to round half *up*, and half *down* in the long run.

2.6 SIGNIFICANT DIGITS

In the foregoing analysis it should be clear to the student that the relative degree of precision in reporting a measure depends upon the number of digits which supply information, irrespective of the position of the decimal point. Persons unaccustomed to computations such as those in statistics think of measurements of precision in terms of the number of decimal points carried, but this obviously depends upon the magnitude of the unit involved. In school expenditures a report of $2,280,000, a rounded figure, has three significant digits—the first 2, the second 2, and the 8. It has the same number of significant digits and hence the same relative precision as 2.28 reported as the millions of dollars spent, or as the average number of trials required for subjects to solve a problem successfully.

The student should develop skill in discerning readily the number of significant digits in figures with which he deals. The following are a series of approximate numbers and the number of significant digits of each:

Number	Number of Significant Digits
5.35	3
535	3
535,000	3
0.000535	3
53,500	3
3.00535	6
2.300	4
2,300.001	7

2.7 ROUNDING IN COMPUTATION

Although by now we should be fairly agreed on rules for rounding a given number, and how to report approximate numbers, we have yet to discuss the question of how to handle approximate numbers in computational work. We may be inclined to reason that, since we often deal with *rounded* or *approximate* numbers and numbers which involve many other types of error, it would be wasteful and inefficient to devote much energy to avoiding computational error. As we gain experience with computations involved in the study of statistics, we find that this attitude leads only to trouble. *We should approach our computational work in the frame of mind that each statistic, each derived figure, is to be reported correct and exact to the number of significant digits which seem required and that no derived figure or statistic is known to be correct and exact until the computations have been checked and verified.*

If at all possible, we should have access to calculating machines so that carrying large numbers of figures in our computations will not be a burden. As a general rule, it should be remembered that errors due to rounding comprise only one class of error we shall encounter. Figures will be copied wrong. They will be added wrong. The right figure will be on the sheet, but the wrong figure will be entered on the keyboard of the calculator, etc. Furthermore, *substantial errors may result from rounding too much.* If in doubt, we should retain a large number of figures in the computations. Retaining digits, as a general rule, is less likely to prove disastrous. As experience is gained with computations, we shall learn to apply some of the rules given below and to set up computation sheets so that we may carry "no more decimals than are necessary." In statistical work involving a great deal of computation, it is important to be able to minimize the number of figures that are to be handled, not only to save time but also to avoid greater opportunities for clerical error.

Experience suggests the advisability of noting one further caution at this point. A distinction must be made between *computational* errors and errors of *measurement*. Owing to lack of reliability, poor standardization, and other matters there may be many types of measurement error in an educational measurement. From the point of view of test unreliability, therefore, we would say that a score of 53 is accurate *within limits*. In the sense of the present discussion, however, the score of 53 would not be considered an "approximation." Objective tests are usually scored in such a way that (barring clerical errors) a given set of responses produces only one "exact" score. A verified scoring in the above hypothetical case is considered to yield the one and only score, 53. It will hence, *not* be considered an approximate number for computational purposes.

(*a*) MULTIPLYING AND DIVIDING APPROXIMATE NUMBERS

In multiplying one approximate number by another or in dividing one approximate number by another, the result (product or quotient) may be carried to the same number of significant figures as the number of significant figures in that one of the two numbers which has the lesser number of significant digits.

Examples:

$$456.5 \times 2.4 = 1{,}100$$

(If both factors are "rounded" or approximate, the figure with the least number of significant digits is 2.4. It has only two significant digits; therefore, although the absolute product is 1,095.60, we round our result to two significant digits.)

$$221 \times 2{,}351 = 519{,}571$$

(*Only if both numbers are exact, involving no rounding, no approximation.* If the first number, 221, is an approximate number, for instance, one that is rounded, the result would be 520,000, the first zero being significant. If the 221 is exact and the 2,351 is an approximate number, we would be entitled to report under this rule 519,600 as the result, using in this case four significant digits.)

$$5.6534 \div .0159 = 356, \text{ not } 355.55597, \text{ etc.}$$

(If both are approximate.)

$$5.1 \div 8.379 = .61, \text{ not } .60866, \text{ etc.}$$

(If both are approximate.)

It is to be noted in the application of this rule that the restriction applies only to *approximate numbers*. That is to say, in multiplying together two numbers or in dividing one number by the other, if both are exact numbers and no approximation is involved, there is no restriction on the number of digits which may be used in the result. The rule as stated applies to the case where both numbers are rounded. In this case, the product or the quotient has as many significant digits as the number with the least number of significant digits. If one of the two numbers, however, is exact and the other approximate, the product or the quotient may have the number of significant digits that the approximate number has.

(*b*) SQUARE ROOT OF AN APPROXIMATE NUMBER

The square root of an approximate number may be reported to as many significant figures as the number itself contains. For example, the square root of 24.56 may be reported as 4.956, that is, to four significant figures. It would be improper to report 4.95580, which contains six significant figures, and unnecessary to round to 4.96. Similarly, the square root of .0135 may be reported as .116 but not as .1161895 or .1162. It is to be noted that the rule for significant digits for the *square of a number* is covered by the rule in (*a*) above, dealing with multiplication. For instance, the square of the approximate number, 79.3, would be reported as 6,290, even though the actual result of squaring the number is 6,288.49. The square of 79.3, of course, is the product of 79.3 and 79.3 two approximate numbers each of which has three significant digits.

(*c*) ADDITION AND SUBTRACTION OF APPROXIMATE NUMBERS

Digits of an algebraic sum to the right of that place in which occurs the last significant digit of any of the numbers summed are not significant. For example, suppose that we are to get the algebraic sum of 321.51, minus 21.6, plus 4.568, and minus 115.25. If these are all approximate numbers, we know that the last digit is the "nearest number" in that place. In arraying the four figures, one under the other, and properly lining up the decimals, we have:

$$\begin{array}{r} 321.510 \\ -21.600 \\ +4.568 \\ -115.250 \\ \hline 189.228 \end{array}$$

In the above, we have put in ciphers to line all the figures up with three decimals just as we would do in entering them in a calculating machine and performing the algebraic addition. When we do this, of course, for the first figure, we have .510 at the right of the decimal, and we know, it being an approximate number, that it is probably not exactly .510, but is somewhere between .505 and .515. It could be exactly .510 only by accident. Similarly, in the second figure, we have written 21.600, but from our rule of rounding, we know this may mean anything from 21.550 to 21.650. Only the third of our series of figures is dependable in the third decimal place.

In the first decimal place to the right occurs the rounded digit for the second number. That is the place in which appears the last significant digit of the second number. No other numbers have last significant digits to the left of that place. Therefore, according to this rule, we may report our algebraic sum as 189.2. The figures which were derived from the actual computation on the calculator to the right of the first decimal

8. What is the effect on σ of adding a constant to each score? Of multiplying by a constant?

9. Which of the diagrams in Chapter 3 would be useful in converting raw scores on a test to *percentile scores*? What proportion of percentile scores of the Navy Recruits would be between 70 and 79, inclusive? 20 and 29, inclusive? What kind of frequency distribution would we have if all scores are converted to percentile scores?

10. With what kind of skewness will scores be distributed on a test made up of (*a*) most items too easy for the group tested? (*b*) Items too difficult for the group?

11. In what respect is a distribution of high-school enrollments similar to a distribution of teachers' salaries.

12. Compute the mean and standard deviation of the 500 scores in Table 3.1. Have some members of the class use 57 as the arbitrary origin, some use 12, and others some other class mark. Compare computations by the three methods. What are the advantages of each?

13. The standard deviation of Test A is 5.6 and its mean is 31.8. The standard deviation of Test B is 7.5 and its mean is 83.9. Using equation 4.16 compute standard scores corresponding to raw scores of 15, 24, and 39 on Test A and 71, 89, and 99 on Test B. What raw scores for Test A and Test B correspond to Z of 64?

14. The standard deviation of mental age in months of group A is 14 and the mean is 104. The standard deviation of group B is 17 and the mean is 193. Which would you consider the more homogeneous group?

REFERENCES

1. Freund, John E., *Modern Elementary Statistics*, New York, Prentice-Hall, 1952, Chapter 5.
2. Guilford, Joy P., *Fundamental Statistics in Psychology and Education*, Second Ed., New York, McGraw-Hill Book Co., 1950, Chapters 5 and 6.
3. Johnson, Palmer O., and Robert W. B. Jackson, *Introduction to Statistical Methods*, New York, Prentice-Hall, 1953, Chapters 4 and 6.
4. Kenney, John F., and E. S. Keeping, *Mathematics of Statistics*, Third Ed., New York, D. Van Nostrand Co., 1954.
5. Lindquist, Everet F., *A First Course in Statistics*, Revised Ed., Boston, Houghton Mifflin Co., 1942, Chapter 6.
6. Walker, Helen M., *Elementary Statistical Methods*, New York, Henry Holt and Co., 1943, Chapters 8 and 10.

CHAPTER 5

Probability and Theoretical Distributions

In previous chapters we discussed frequency distributions made up from actual data. One of the chief values of the study of statistics is to learn about some of the *theoretical* frequency distributions which serve as models for the interpretation of many types of actual distributions.

Most of the frequency tables of Chapters 3 and 4 were in reality *samples* of large populations. Our interest, however, is seldom in the sample itself, but in the population which it represents. Fundamental to statistical work are the inductive processes of generalizing to a population from a single specific set of observations. Where samples are drawn under conditions which yield only *chance* variations from sample to sample, mathematical models may be used to make probability statements regarding the relationship of a sample to its parent population. The foundation of the mathematical theory is probability. A knowledge of elementary "laws of chance" is, therefore, essential to an understanding of even the most rudimentary principles of statistical theory.

5.1 PROBABILITY AND GAMES OF CHANCE

Although the theory of probability had its beginning in games of chance, the principles apply to many statistical events or occurrences having to do with education. The tossing of a coin will yield either a head or a tail. This is analogous to a single event such as arbitrarily marking a true-false item, selecting from a population of half boys and half girls one individual who may be of either sex, or sampling an individual, in a community poll on a school bond issue, from among those who are "for" and those who are "against." As we shall see, the latter example is different in that the probabilities of the various outcomes may not be known in advance.

In tossing the penny it seems reasonable to us that it will fall heads or tails about the same number of times in many throws. We say that the events H, heads, and T, tails, are *equally likely*. Similarly we would expect, *in the long run*, throws of a single die to stop with one side up about as many times as another. Therefore, the six faces are said to be *equally likely* to turn up in a single throw.

Two or more events are said to be *mutually exclusive* if only one can occur in a single happening. Thus, drawing an ace and drawing a spade are not mutually exclusive because the happening of one *does not* exclude the happening of the other.

5.2 DEFINITION OF PROBABILITY

By the common textbook definition, the probability of an event is the ratio of the number of ways in which it *can* occur to all possible ways. More specifically, the classical or *a priori* definition of probability is as follows: *If an event can happen in s ways out of a total of n mutually exclusive and equally likely ways, the probability of the event is s/n.*

In tossing a coin the occurrence of a head is but one of the two ways the coin can fall. Hence the probability of a head is 1/2 or .5. There are six faces on a single die. The probability of throwing a "four" is but one of the six. It is, therefore, 1/6. As a further illustration, consider the probability of rolling a "seven" with two dice. Since for each of the six ways the first die can fall, the second can fall six ways, $n = 6 \times 6 = 36$. However, for each of the six ways the first die can fall, there is only one way the second can fall such that the total will be "seven." Hence, $s = 6$. Therefore the probability of a seven is $6/36 = 1/6$.

This is not an altogether satisfactory definition. For that matter there does not seem to be an adequate definition of probability. Notice that we have used the term "equally likely" in the sense of "equally probable." The definition is thus circular since the term being defined is used in the definition. Moreover, the definition does not tell us how to know when the ways are equally likely or what the probability is when they are not equally likely. The probability of throwing a "four" with a loaded die may be more than the ideal 1/6 or less than the ideal 1/6.

The statistician gets around such difficulties by using an *empirical* definition. He defines the probability of the occurrence of an event in terms of *relative frequency* in a given population. The probability that a "four" will turn up in throwing the die is the ratio of the number of successes, f, to the number of throws, N, where N is a large population of throws of the die. In an imaginary large number of throws of the unbiased die, we would *expect in the long run* that the ratio, f/N, would be 1/6. The probability of throwing a "four" with the *loaded* die we might not be able to determine exactly, but we could throw it many times and estimate it by computing f/N. Though the probability of a "four" might not be 1/6, we assume that some value, P, exists as a limit approached by the frequency ratio, f/N, as N becomes increasingly large.

This kind of *relative frequency* definition of probability permits direct observation of it in the reality of actual objects or events. Statistics, after all, is a branch of applied mathematics which deals with the real world. We are interested in *a priori* probabilities only because they help us understand the basic ideas behind some of the theoretical distributions which the statistician uses in his interpretations of observations from the real world. There are not many "events" in educational work for which probabilities may be developed *a priori* as in tossing of coins and dice. However, the laws of theoretical *a priori* probability have their counterparts in *empirical* probability. We may thus apply the theory to empirical probabilities derived from experience. A high-school record system, for instance, might show that only 65 of 100 entering freshman graduate. From this information we may estimate the probability, P, of graduation of an entering freshman to be .65. Similarly the probability of drawing a high school of less than 100 enrollment from a card file of all high schools in the United States may be estimated from the table in Section 3.1 to be 9,565/24,314. The experience of life insurance companies is used to compute the chance of a person of a given age living through the next year, the *empirical probability* of living. We are interested in probability theory because it is useful in dealing with such probabilities. As a step in the development of this notion we must first examine two of the classical *laws of probability*.

(*a*) THE ADDITION THEOREM. *The probability that one of a set of two or more mutually exclusive events will happen in a single trial is the sum of the probabilities of the separate events.* In throwing a single die, the probability of *either a* 2 *or a* 3 is $1/6 + 1/6 = 1/3$. In selecting one person at random from a group of 4 freshmen, 6 sophomores, and 8 juniors, the probability of selecting *either a sophomore or a junior* is $6/18 + 8/18 = 7/9$. The probability of getting either a freshman, a sophomore, or a junior is $4/18 + 6/18 + 8/18 = 18/18 = 1$ or certainty.

(*b*) THE MULTIPLICATION THEOREM. *The probability that two or more of a set of mutually independent events will all happen is the product of their probabilities.* The events are said to be mutually independent if the happening of one is in no way affected by the happening of the other. In tossing two different (independent) coins, the probability that one comes up heads is 1/2, and the probability that the other comes up heads is also

1/2. The probability that *both* will come up heads is, therefore, (1/2)(1/2) = 1/4.

The probability that the first of two dice will be 2, and *once that has occurred* the second will be 3, is (1/6)(1/6) = 1/36. However, if the 2 and 3 need not fall in that order, the probability is 2/6 that the first die would show either 2 or 3, and *once that occurred* the probability is 1/6 that the second die would come up with whichever one of the two faces, 2 or 3, would be required. Thus (2/6)(1/6) = 1/18 would be the probability of a 2 and a 3. This may be verified by the addition theorem since the 2–3 occurrence we saw had a probability of 1/36, and similarly the 3–2 occurrence has a probability of 1/36. The probability that *either* would occur would be 1/36 + 1/36 = 1/18.

We are now ready to experiment with the combining of probabilities and discover some mathematical devices of great convenience in practical statistical work.

5.3 AN EXPERIMENT IN TEACHER JUDGMENT

Let us assume that an investigator has some definite hypotheses regarding abilities of teachers to predict observable behavior of pupils. For our discussion the behavior observed may be almost anything—objectively measurable responses on a paper and pencil test; preferences or alternative choices among ideas, opinions, or objects; or success in the completion of some task such as solving a puzzle or an arithmetic problem or clearing a 4-foot high jump. To be specific, we will assume that the teacher has been asked to predict which of three pupils will solve an arithmetic problem correctly. The pupils are given the problem, and it is found that of the three pupils, *a*, *b*, and *c*, pupil *b* correctly solved the problem. The goal of the investigator is to evaluate the results of the teacher's judgments in the light of what might be expected to *happenquite by accident*.

The above theorems on probability assist us in assessing the possibilities. We may use the coin-tossing model for this purpose. Suppose that the teacher possessed no ability whatsoever to predict the outcome of the experiment. Imagine that he would, in the long run, predict the outcome correctly only about as often as he would predict it incorrectly. Then suppose that he were to predict solutions by the three pupils. The process would then be analogous to tossing three coins for which, let us say, a head, *H*, would be comparable to a *correct* judgment, and tails, *T*, an *incorrect* one. The possible events are as follows:

$$\begin{array}{cccc} HHH & HTH & THH & TTT \\ HHT & HTT & THT & TTH \end{array}$$

By means of the multiplication theorem we could directly compute the probability of tossing three heads (making three correct judgments), $(1/2)(1/2)(1/2) = 1/8$, if the probability is 1/2 for each of the three independent events. This checks with our observation that there is only one of the eight arrangements that consists of *all heads*. We now note that there are three ways in which the teacher could (quite by accident) get two of the three correct, *HHT*, *HTH*, and *THH*. Proceeding in this manner, we may show that the teacher's chances for various possible "scores" of correct judgments are as follows:

Number Correct (Heads), X	Number of Ways, f	Probability ($P(X)$)
3	1	1/8
2	3	3/8
1	3	3/8
0	1	1/8
Total	8	8/8

In this tabulation we use X to represent the various *scores*, or correct judgments, f the number of ways each X may happen out of 8, and $P(X)$ the probability of X.

It should be recognized at once that we have here a *theoretical* frequency distribution—theoretical in the sense that, *in the long run*, and if chance alone were operating, we would expect to get frequencies for the different scores in proportion to 1, 3, 3, and 1, respectively.

It is important to note that the values in the last column show for each score, the probability, $P(X)$, of that score, happening by chance. It is noted further that the 4 probabilities sum to 1. Such distributions of *relative frequencies* or *probabilities* will be used frequently throughout the study of statistics. An important point to remember is that $\Sigma P(X) = 1$.

In Section 3.2 we used the histogram to represent frequency distributions. The same device may be used to represent theoretical distributions. The theoretical distribution of the teacher's judgments is pictured as a histogram in Fig. 5.1. The outcome, X, can be only a *discrete count* of successes. It is not a measure on a continuous scale. Therefore, a value of X, such as 2, cannot represent a range of possible values, that is, 1.5 to 2.5. There are advantages, however, in representing distributions as *area*. We therefore construct the histogram with columns centered at each possible value of X, with width of 1, and height equal to the corresponding $P(X)$, or relative frequency. Areas now represent probabilities. The sum of the areas of all rectangles in the histogram is 1.00.

There is another major conclusion to be reached from the above experiment: our mathematical model tells us what to expect if an infinitely large number of trials (throws of three coins) were to be made. Information from the above table, not dependent upon any single trial or set of trials, is called a *parameter* as distinct from comparable information from a *sample*, which is called a *statistic*. For instance, in 8 tosses of three coins, the actual relative frequencies (f/N) might be 2/8, 2/8, 3/8, and 1/8.

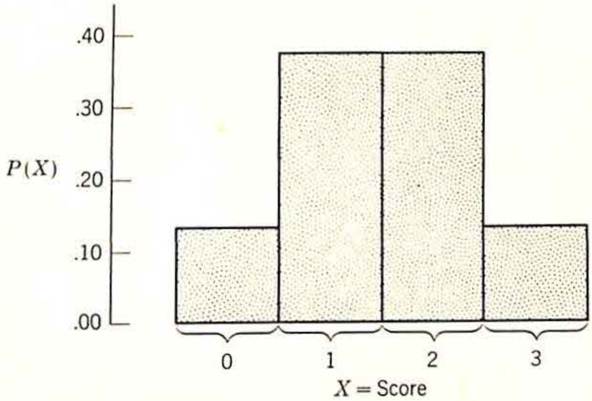


FIG. 5.1. Distribution of probabilities of number of heads in tossing three coins.

Our experiment is an introduction to the type of decision a research worker must make in drawing conclusions from a sample. As we have indicated above, our model reveals what might be expected had the teacher really possessed no skill whatsoever in judging his pupils. In the language of statistics we say that we have set up a *null hypothesis*, the hypothesis that the teacher's decisions are no different from results which could have been derived from some random process such as the tossing of three perfectly balanced coins.

Suppose that, in conducting our experiment, the teacher makes a score of 2, that is, predicts two out of three of the pupil's responses. Reference to our table will show that the probability of 2 *or better*, is .50. Probabilities such as this may be expressed symbolically, using signs for inequalities. The symbols, $<$ and $>$, mean "is less than" and "is greater than," respectively. In combination with the equal sign the former becomes $\leqq$ or $\leq$. This is interpreted as "is equal to or less than."

Similarly the symbols $\geqq$ or $\geq$ are used for "is equal to or greater than." The statement that the probability of a value of X of 2 or more is equal to .50 may be expressed simply as $P(X \geq 2) = .50$.

The above probability means that there is a fifty-fifty chance that *chance alone* would produce a result as good as or even better than that found in the experiment. Even a perfect score of three correct judgments would happen according to our model .125 or 12 1/2 per cent of the time in repeated trials. This too seems so plausible that we cannot reject the hypothesis. As has already been suspected, the experiment is not sensitive enough to enable us to distinguish between *what could just happen* and what the teacher can predict. We shall, therefore, redesign the experiment.

5.4 THE BINOMIAL DISTRIBUTION

If we have the teacher predict the success or failure of 10 pupils, instead of 3 pupils, we could proceed as above to calculate a theoretical distribution showing $P(X)$ for each possible number of correct predictions, $X = 0, 1, 2, 3, \cdots, 10$. Little work at this step will be required to show us that tools are needed to reduce the tedious work involved. Where there were eight arrangements of the three coins in the above table, our new table would have to take into account 1,024 arrangements of 10 contingencies. Fortunately we have an important tool in the *binomial* distribution, sometimes called the Bernoullian distribution, after Jacques Bernoulli who first studied it (in *Ars Conjectandi*, published in 1713).

Suppose that p represents the probability that some event will happen in a single trial. In our coin tossing this would be $p = 1/2$ for a head to occur; in throwing a die, $p = 1/6$ for an ace to occur; or it may be some unknown probability to which a numerical value cannot be assigned *a priori*. Then, if $q = (1 - p)$ is the probability that it *will not* happen in a single trial, the probability of its happening exactly X times in N trials may be shown to be $\left(\frac{N}{X}\right) p^X q^{N-X}$. The symbol $\left(\frac{N}{X}\right)$, sometimes written, ${}_NC_X$ and in other ways, is the number of combinations of N things taken X at a time. From the study of permutations and combinations in college algebra we may write

$$\left(\frac{N}{X}\right) p^X q^{N-X} = \frac{N!}{X!(N-X)!} p^X q^{N-X} \tag{5.1}$$

where $N!$ is factorial N, the product of N and all the positive integers smaller than it. For example, $5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120$.

Thus, the probability of exactly 3 heads in 10 tosses of a coin is

$$P(3) = \frac{10!}{3!(10-3)!}\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^7 = \frac{120}{1,024} = .1172$$

In a similar manner we may actually compute from equation 5.1 the other theoretical relative frequencies for our experiment as revised.[1] These are shown in Table 5.1 and the histogram of Fig. 5.2.

We see in Table 5.1 that if we were to engage in many chance events such as the tossing of ten coins, *in the long run*, all ten would be "successes" once in 1,024 times, or $P = .00098$. Similarly we see that $P(9) = .00977$, $P(8) = .04395$. Combining these probabilities by the addition theorem, we would expect a 9 or 10 about 1.1 percent of the time; an 8, 9, or 10 about $5\frac{1}{2}$ percent of the time.

Now suppose that we decided in advance that we would *reject* the null hypothesis (that the teacher's success in prediction is very much like tossing a coin) in case the probability of occurrence of the results of the experiment simply by chance is *so low* that the assumption of random occurrence seems unreasonable. Let us say that we agree that we will reject the hypothesis if $P \leq .05$, that is, if the probability of the observed result is *equal to or less than* 5 percent. We are then willing to run the risk of .05, 5 chances in 100, of being wrong in ascribing some ability to the teacher's judgment when a great many repetitions of the experiment might in fact show that his decisions were strictly random.

On completion of the teacher's judgments we find that he has correctly predicted eight of the ten pupils. As we have noted above, the probability of this result on the null hypothesis is $P(X \geqq 8) = .0547$, that is, by chance alone 8 or better (an 8, 9, or 10) could be expected 5.47 percent of the time. Therefore, we *do not* reject the null hypothesis at the 5 percent level.

Notice that we have not proved that the teacher does not have genuine competence in understanding the abilities of his pupils. We have simply said that from this experiment we are not ready to classify the result as something probably distinct from chance occurrence.

There are several features of Table 5.1 to be emphasized. In the first place, the distribution is symmetrical as may be seen from the last two columns. This is true of binomial distributions only for which $p = 1/2$.

[1] Other computing aids may be used to determine these probabilities. Tables for the binomial expansion may be found for values of N up to 50 in Churchill Eisenhart, editor, *Tables of the Binomial Probability Distribution*, Applied Mathematics Series, No. 6, National Bureau of Standards, Washington, D. C., 1950. Rarely, however, is it necessary to compute these probabilities. As we shall see, there are other distributions which closely approximate the binomial.

TABLE 5.1

THEORETICAL BINOMIAL FREQUENCIES, N = 10; $p = 1/2$

Number of Successes, X	Number of Ways (in 1,024) f	Probability, $P(X)$
10	1	.00098
9	10	.00977
8	45	.04395
7	120	.11719
6	210	.20508
5	252	.24609
4	210	.20508
3	120	.11719
2	45	.04395
1	10	.00977
0	1	.00098
Total	1,024	1.00003

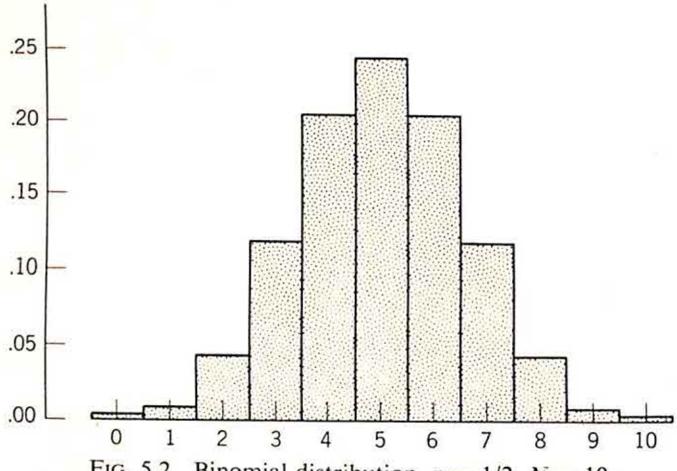


FIG. 5.2. Binomial distribution, $p = 1/2$, $N = 10$.

In evaluating expected frequencies or probabilities of an ace in five throws of one die, p would be 1/6. From equation 5.1 we would find the distribution for no aces, one ace, two aces, etc., to be in proportion to: 3,125, 1,250, 250, 25, and 1. This distribution is decidedly skewed in the positive direction. Binomial probabilities thus depend upon the value of p (or q) and the number of trials, N. For each combination of N and p, therefore, there is a distinct set of binomial frequencies.

The application of equation 5.1 to determine binomial probabilities for successive values of X is equivalent to finding the terms of the binomial expansion

$$(q+p)^N = q^N + Nq^{N-1}p + \frac{N(N-1)}{2}q^{N-2}p^2 + \frac{N(N-1)(N-2)}{(2)(3)}q^{N-3}p^3 + \cdots + p^N \qquad (5.2)$$

A general representation of tables like Table 5.1 appears as Table 5.2.

TABLE 5.2

THE BINOMIAL DISTRIBUTION

Score, X	Relative Frequency, $P(X)$	Weighted Score, $XP(X)$
(1)	(2)	(3)
0	q^N	0
1	Npq^{N-1}	Npq^{N-1}
2	$\frac{N(N-1)}{(1)(2)}p^2q^{N-2}$	$N(N-1)p^2q^{N-2}$
.	.	.
.	.	.
.	.	.
X	$\frac{N(N-1)\cdots(N-X+1)p^Xq^{N-X}}{X!}$	$\frac{N(N-1)\cdots(N-X+1)p^Xq^{N-X}}{(X-1)!}$
.	.	.
.	.	.
.	.	.
N	p^N	Np^N
Total	$\Sigma P(X) = (q+p)^N = 1$	$\Sigma XP(X) = Np(q+p)^{N-1} = Np$

Before using it to make some general statements about the binomial we will first work with the actual distribution in Table 5.1. Table 5.1 may be treated just as we have treated other distributions in Chapters 3 and 4. The mean number of successes, $\bar{X}$, may be found from the first and second columns from $(\Sigma fX)/N$. This would be the same as $\Sigma(f/N)(X)$. That is,

$$\bar{X} = \frac{f_1X_1 + f_2X_2 + f_3X_3 + \cdots}{N}$$

or

$$\bar{X} = (f_1/N)(X_1) + (f_2/N)(X_2) + (f_3/N)(X_3) + \cdots \tag{5.3}$$

The expression f_i/N is simply the relative frequency or, in the theoretical distribution, the probability $P(X)$. We may thus use the third column of Table 5.1 to compute the mean of the distribution by

$$\bar{X} = \Sigma XP(X) \tag{5.4}$$

Cumulative multiplication on a computing machine should verify the result in Table 5.1 of $\bar{X} = 5.0$. A glance at the histogram (Fig. 5.2) confirms this *central value* of the distribution.

The foregoing analysis leads to a concept important in statistical theory. Note that the kind of operation represented by our theoretical model of Table 5.1 and Fig. 5.2 is purely a chance, or *random* operation. The mean which we obtained from equation 5.4 is the theoretical mean which you would *expect* to get *in the long run* by repeating the operation (coin tossing or what not) many, many times. The *expected value*, $E(X)$, of a variable is the mean of its probability distribution. Hence from equation 5.4

$$E(X) = X_1P(X_1) + X_2P(X_2) + \cdots + X_NP(X_N) \tag{5.5}$$

In other words, the mean value, or expected value, $E(X)$, of a variable, X, from a probability distribution, is the sum of all the values that X can take, each weighted by its probability or proportion, $P(X)$.

By the same line of reasoning, the expected value, $E(X^2)$, of the square of X is the mean of its probability distribution. That is, as in equation 5.5,

$$E(X^2) = \Sigma(f/N)(X^2) = \Sigma X^2P(X) \tag{5.6}$$

Using the calculator, we find the cumulative product of $P(X)$ and the squares of X in Table 5.1 to be

$$E(X^2) = 27.5$$

According to equation 4.14, the variance is the "mean of the squares minus the square of the mean."

$$\sigma^2 = (\Sigma fX^2)/N - (\Sigma fX/N)^2$$

From equations 5.5 and 5.6, the equivalent definition of variance in a theoretical distribution is

$$\sigma^2 = E(X)^2 - (EX)^2 \tag{5.7}$$

Having found that in Table 5.1 $E(X^2) = 27.5$, and $E(X) = 5.0$, we may compute

$$\sigma^2 = 27.5 - 25.0 = 2.5$$

The square root of this, 1.58, is the standard deviation of this theoretical distribution.

We should now be able to apply this approach to Table 5.2 to discover something about binomial distributions in general.

5.5 THE MEAN AND VARIANCE OF BINOMIAL DISTRIBUTIONS

From equations 5.4 and 5.5 we may find an expression of the mean, μ, of any binomial. We are using the Greek letter mu, following a convention common to statistics of using Greek letters for a *parameter*, that is, some measure derived from a total universe or an *ideal* or theoretical population. We might have introduced this symbolism in connection with equation 5.4, but we chose to stay with the notation, $\bar{X}$, with which we had already become familiar. But since Table 5.1 is a binomial distribution, not merely a sample from such a distribution, μ might more appropriately have been used. In the future we will reserve $\bar{X}$ to represent only the mean from an actual *sample* of values from some population.

Referring to column 3 of Table 5.2, we find, for each value of X, (0, 1, 2, etc.) corresponding values of $XP(X)$. The terms of the binomial expansion, it will be recalled, are the *relative frequencies*, $P(X)$. They appear as column 2. From column 3, a common factor Np may be extracted, leaving a series of terms which are just the terms for the binomial $(q + p)^{N-1}$. This binomial is equal to unity, because $q + p = 1.00$, and 1.00 raised to any power is 1.00. Hence $\Sigma XP(X) = Np$. Thus the mean of a binomial distribution is

$$\mu = E(X) = Np \tag{5.8}$$

By a similar, but somewhat more complicated process, we could sum products of column 1 and column 3 to get $\Sigma X^2P(X)$ in terms of the binomial expansion. We would find that it reduces to $E(X^2) = Np\,[(N - 1)p + 1]$. From this we must subtract the square of the mean, $\mu^2 = N^2p^2$. Whence, from equation 5.7,

$$\begin{aligned} \sigma^2 &= Np\,[(N-1)p + 1] - N^2p^2 \\ &= Np - Np^2 \\ &= N(p - p^2) = Np(1 - p) \\ &= Npq \end{aligned} \tag{5.9}$$

There is a system (the moment system) which uses the expected values of X^3 and X^4 to develop measures of skewness and peakedness of theoretical distributions. Because development of these measures is not essential to our objectives here, it will be omitted. The results, however, establish some general principles about binomial distributions which should be known.

The measure of *skewness* for the binomial is

$$\gamma_1 = \frac{q - p}{\sqrt{Npq}} \tag{5.10}$$

When $\gamma_1 = 0$, the binomial is symmetrical. When γ_1 is positive, the binomial is positively skewed; when negative, negatively skewed.

The measure of peakedness or *kurtosis* is

$$\gamma_2 = \frac{1 - 6pq}{Npq} \tag{5.11}$$

When $p = q = .5$, $\gamma_2 = -2/N$. In this case the binomial is *moderately* convex or approximately *mesokurtic* except for very small N, say less than 10. Values of γ_2 larger than this, particularly large positive values, indicate so-called leptokurtic distributions; smaller values (that is, negative values) of γ_2 indicate platykurtic distributions. These are often more peaked or flatter, respectively, than the normal distribution, but contrary examples are known.

Equations 5.10 and 5.11 indicate the following conclusions concerning the binomial distribution:

(*a*) The numerator in equation 5.10 approaches zero as p and q approach the value 1/2. Hence skewness is zero when $p = q = 1/2$.

(*b*) The fraction, γ_1, becomes smaller as N increases. Hence skewness in the binomial becomes negligible as the number of terms in the binomial becomes larger and larger.

(*c*) As N increases, the fraction in γ_2 becomes negligible and γ_2 approximates 0. Thus with large N the kurtosis in the binomial is negligible.

Figure 5.3 is the histogram of a binomial distribution, with $p = .3$, $q = .7$, and $N = 20$. Even though p and q differ considerably from 1/2, with $N = 20$ the distribution is reasonably symmetrical, and it is mesokurtic as would be expected from conclusion *c* above. From equation 5.10 the measure of skewness is $\gamma_1 = .195$. A clear case of skewness easily observed from a histogram or frequency polygon would have a γ_1 either positive or negative of at least .4.

From equation 5.11 we find $\gamma_2 = -.06$. The mesokurtic or "normal" value of γ_2 would be $-2/20 = -.10$. The distribution of Fig. 5.3 may therefore be considered approximately "normal" as to peakedness or kurtosis. As a matter of fact, one object of this chapter has been to show how the binomial distribution is related to another very important theoretical frequency distribution, the so-called normal distribution.

We will see in Chapter 6 that this characteristic of the binomial to approach symmetry and mesokurtosis with large N is one reason why it may be approximated by the normal curve, thus saving much computation.

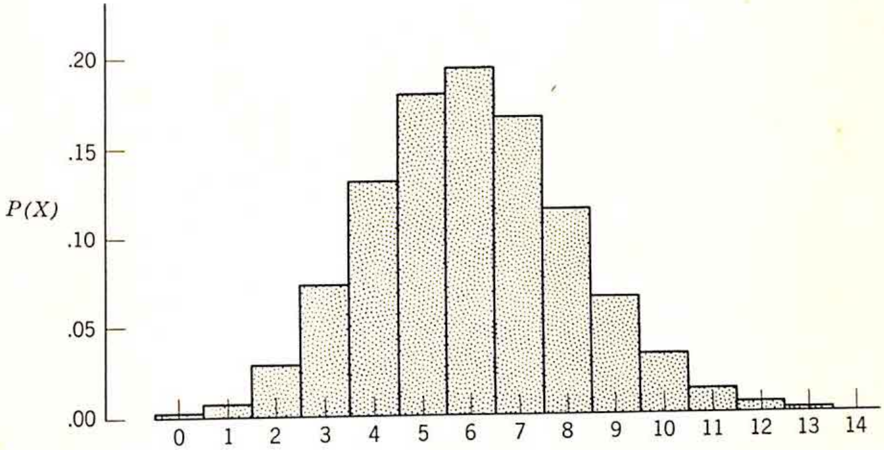


FIG. 5.3. Binomial distribution, $p = 0{\cdot}3$, $N = 20$.

As we have noted, the binomial is a "discrete" distribution, that is, one which includes only integral values. For this reason it is sometimes called the *point binomial*. In the next chapter it will be approximated by a smooth curve or *continuous* distribution, one for which individual measurements may take on any value within the limits of the distribution.

5.6 SAMPLE AND UNIVERSE

To this point in the development of this chapter we should have been building up some notions of the role played by sampling in statistics. A review of the major ideas of this chapter should help us establish a few of the understandings and attitudes basic to most statistical work.

We have used the terms *population* and *universe* synonymously, as a collection of individuals, objects, or events having some common measurable characteristic. The population is the entire set of individuals or measurements regarding which we wish some information. A population may thus be the salaries of teachers in the United States, or it might be specified as only the population of salaries of (1) elementary school teachers, (2) teachers with less than four years of training beyond high school, (3) teachers in the State of Missouri, (4) teachers employed in public schools, (5) those teaching during the year 1952. Our interest might be with a theoretical population of number of heads based upon hypothetically tossing ten coins an infinite number of times, or it might be a practical population of the hypothetically infinite number of judgments a teacher could make of pupils, or of tasks or test items that a pupil could conceivably be asked to perform. The population may be *finite*, such as achievement scores of the 28 students in ninth-grade English class, on the one hand, or, on the other, so large—such as the heights of all 12-year-old boys in the United States—that for practical purposes it may be considered *infinite*.

In great part the contribution of statistics is in securing information about a population by generalizing about it from a part of it called a *sample*. The characteristic of the population is called a *parameter*, for example, mean, median, variance, proportion, percentile. The corresponding measure from the sample is called a *statistic*. In the day-to-day operations of a school, teachers and administrators observe *statistics* from a subset of observations or measurements, and from these draw *inductive inferences* about the *parameters* of the entire class of all similar such observations or measurements. Teachers predict from samples, good ones well and poor ones poorly, the difficulties certain types of pupils will have in learning a unit of instruction, the "amount of learning" of individual pupils (from tests made up of samples from an infinite number and variety of exercises which might be taken to measure successful learning), or the effectiveness of a teaching method or instructional device (from having tried out or conducted an experiment with it in a sample of one or more classes). A principal or superintendent predicts costs, enrollments, absences, drop-outs, opinions of the teaching staff, the school board, or the public, or the superiority of a teacher with one set of characteristics over one with another set of characteristics. A good principal predicts well, a poor one poorly. Similarly the educational researcher generalizes from the results of one or more experiments to the entire universe of such experiments.

In statistics we call this process *statistical inference*. The statistically literate educationist knows that the results of such inductive reasoning are not exact. He knows that such inferences may be made only within a realm of uncertainty, which under proper circumstances may be measured in terms of probability.

In this chapter we have studied the elements of probability theory as (*a*) an introduction to the role of probability in statistical inference and (*b*) the basis for developing some understanding of the binomial distribution. We have found the binomial distribution directly useful in making probability statements concerning *statistical hypotheses* about a population parameter. Several later chapters will deal largely with ways of testing statistical hypotheses. For the most part the testing will be done with mathematical models or probability distributions, as was done with the binomial distributions in this chapter. One reason for our study of the binomial is that it leads to one of these distributions, a continuous curve or *distribution function*, which we will examine in the next chapter.

We have been introduced in this chapter to a new concept in connection with moments of the binomial, the concept of expected value, $E(X)$, as an average value of a variable or of a function of a variable. It is a kind of theoretical average indicating what we might approximate in the long run. We do not actually "expect" to get it. For instance, the probability, p, of a pair in rolling two dice is $1/6 = .167$. In ten throws of two dice the expected value is $E(X) = Np = 1.67$. The sample value may be only a discrete value $(0, 1, 2, 3, \cdots, 10)$ and could thus never be 1.67. Even in many repeated samples of this operation we would not expect an average of exactly 10/6. The following are the results of two sets of ten samples each of ten throws of two dice, where X is the number of pairs:

X	0	1	2	3	4	5	6
Set I	1	5	4	–	–	–	–
Set II	1	3	3	2	–	–	1

The means of the two sets are respectively 1.3 and 2.1, values which are reasonably likely to occur, though the "expected" value may be 1.67.

5.7 CHARACTERISTICS OF A SAMPLE

In this chapter we have been writing about samples, but we have not defined a sample other than to imply that it is a subgroup of measures from a population of measures. In common usage the term *sample* is taken to mean any portion or specimen of the whole. In statistics, however, it usually has a more restricted meaning. In the first place, to be a useful sample, it generally must have an element of *randomness*. Otherwise it is not possible to make a statistical inference or test a statistical hypotheses as we have defined it here. Therefore, throughout this book, the term *sample* will mean *random sample*, except where otherwise specified. Sometimes the investigator is confronted with a sample which is not random, but he does not know it and so applies a statistical test the result of which is invalid because the assumption of randomness inheres in such tests.

A random sample is made up of individuals, or elements, selected in such a way that:

(*a*) All individuals in the population have an equal chance of being included.

(*b*) The drawings of the elements in the sample are independent.

It will be recognized that these are the identical conditions upon which the binomial theory was based. Suppose that one of two dice involved in our experiment of Section 5.6 was loaded so that "six" would come up more frequently than other values. Then computations from the binomial distribution of μ and σ^2 using $p = 1/6$ would not be applicable.

Sometimes a small group may be subjected to sampling for the purpose of establishing the nature of that small group itself, not regarding it as a sample of a large population. The purpose of sampling would be to reduce the work involved compared to taking the whole group into consideration. When a sample of several individuals is chosen at random, but so that each individual can come into the sample only once, we have sampling *without replacement.* The finite size of the population being studied affects various statistics derived from such samples. For example, a sample mean may be expected to depart less from the true mean of the entire group when the sample comes from a finite population than it would if it is of the same size but comes from a similar but infinite population.

Sampling without replacement has a progressively smaller effect the larger the population. Ways of dealing with such samples from small populations are given in Chapter 7.

A questionnaire sent to a sample of parents of students in the senior high school of a community would not be a random sample of the adults in the community, since adults without children or without children in the high school would have zero probability of being chosen, and parents with more than one child in the high school might have a greater probability of being included than those with only one.

Sometimes the term *representative* is applied to sampling. Frequently this is taken to mean, for example, a hand-picked group of students selected by someone well acquainted with them who could therefore pick "typical" individuals. This is *not* a random sample. Often the term is used in connection with a group selected in such a way that there are a given number of individuals of each of several categories of age, economic level, etc., and such that on these variables the group will be a sort of *miniature* of the universe. Only if within each such category, or stratum, the selection is random can a mathematical model be developed which will yield a probability statement regarding the sample in relation to its population. Sometimes nonrandom samples of this type (*cross section samples*, *quota samples*, etc.) are used where it is not convenient to select individuals on a random basis.

One noted expert on sampling in statistics makes a distinction between the *probability sample* and the *judgment sample*.[1] The former corresponds to the definition which we intend to use and the latter to samples chosen by judgment rather than by the automatic methods of random selection. Either type is distinguished from the "chunk," that is, some convenient part of a whole, neither randomly selected nor "representative" of the population.[2]

Obviously, one of the first considerations in sampling is the definition of the population for which generalization is to be made. An experiment conducted in a university laboratory school made up predominantly of pupils above average in intelligence from middle-class homes of an academic community, taught by superior teachers, would yield results which could not logically be claimed to apply to a population of all school children in the region of the same age and grade level.

We will learn more and more about sampling and statistical inference as we take up further topics in statistics. At this point we should have a fair notion of the meaning of *sample* in statistics and should understand the meaning of the term *random* as we will use it in contrast with the less rigorous ideas of *haphazard* or *hit-or-miss*. In Appendix B will be found a table of random numbers, a convenient aid in sampling. Its use will be explained in a later chapter.

EXERCISES

1. Seven out of ten randomly selected voters in a school district are found to be in favor of building a new high-school building. It has been claimed by a local organization that only 40 percent ($p = .40$) of the voters are in favor of the new building. What is the probability that a sample observation of seven or better favorable voters in a sample of ten could have occurred by chance from a population in which $p = .4$?

2. If three cards are drawn from a deck of 52 cards, what is the probability that the first will be an ace, the second a king, and the third a queen?

3. Over a period of several years it has been found that 40 percent of the students in a high-school geometry class *fail* a given test item at the end of the term. What is the

[1] See W. Edwards Deming, *Some Theory of Sampling*, John Wiley and Sons, 1950.

[2] Was it education which someone once called a science of sophomores because so much research reported in education is based on experiments with the author's education classes? Certainly such "samples of convenience" should be important educationally. "Pretesting" of hypotheses and tryout of techniques should be excellent laboratory material for students of classes. The difficulty is with the generalization.

probability that of a random sample of ten students examined at the end of the term (*a*) all ten will *pass* the item, (*b*) nine will pass, (*c*) not more than two will fail?

4. Define:

Probability
Random
Sampling
Statistical inference
Statistical test
Statistical hypothesis
Null hypothesis
Point binomial
Expected value
Parameter
Statistic

5. A baseball player has a batting average of .300. What is the probability of his getting five hits out of six times at bat?

6. The percentage of A grades in educational psychology in a teachers college averaged 20 over several years. What is the probability that five of a random sample of ten students will receive A?

7. What is the probability of at least two heads in six tosses of a coin? At most two heads?

8. What is the expected value of the sum of the values in a throw of two dice?

9. In drawing five cards from a deck of 52 (without replacement), what are the chances of drawing all four aces?

10. What are the conditions of randomness in sampling?

11. By random methods 100 women are selected from a telephone directory in a middle-sized city. Comment on this procedure as a means of selecting a sample for polling consumers in the community.

12. Suppose that we are to test the hypothesis that the proportion of 12-year-old school children from families whose principal wage earners are classified in certain occupational categories is .30. Suppose further that a random sample of 20 such children is to be drawn and interviews in homes are to supply information for testing the hypothesis. We designate the number of households in the sample with the specified characteristic as X. Then, since $p = .3$ and $N = 20$, probabilities of X under the hypothesis are the binomial probabilities in Table 6.1. If the hypothesis is true:

(1) Describe in words and find the value of

(*a*) $P(X > 13)$
(*b*) $P(X < 4)$
(*c*) $P(X \leq 3)$
(*d*) $P(X = 13)$
(*e*) $P(4 < X < 8)$
(*f*) $P(4 \leq X \leq 8)$
(*g*) $P(X \geq 11) + P(X \leq 2)$
(*h*) $P(X > 12) + P(X < 3)$

(2) What is the probability that X is:

(*a*) Greater than 12?
(*b*) At least as much as 10?
(*c*) Not less than 3?
(d) Three or less?
(*e*) Between 3 and 12 inclusive?
(*f*) More than 12 or less than?

(3) Write equations expressing your answers in (2).

13. In a high school of 8 teachers there are 40 different courses to be taught. Assuming each teacher is to teach 5 courses, how many combinations of teaching assignment may be made of the 8 teachers to handle the 40 courses?

14. In a given population it is reasonably tenable to assume that a person's height and his intelligence are independent. On the basis of this assumption, what is the

probability that a person is both "tall" and "intelligent" in a population which, from measurements with certain standards of definition, has been shown to consist of 12 percent "intelligent" persons and 17 percent "tall" persons?

15. A sample of students in a school is drawn by selecting every tenth card in an alphabetical file of all students. Is this a simple random sample?

REFERENCES

1. Deming, William E., *Some Theory of Sampling*, New York, John Wiley and Sons, 1950, Chapter 1.
2. Edwards, Allen L., *Experimental Design in Psychological Research*, New York, Rinehart and Co., 1951, Chapter 3.
3. Eisenhart, Churchill (editor), *Tables of the Binomial Probability Distribution*, National Bureau of Standards, Applied Mathematics Series 6, Washington, D.C., Superintendent of Documents, Government Printing Office, 1950.
4. Freund, John E., *Modern Elementary Statistics*, New York, Prentice-Hall, 1952, Chapter 7.
5. Hansen, Morris H., William N. Hurwitz, and William G. Madow, *Sample Survey Methods and Theory*, New York, John Wiley and Sons, 1953, Vol. 1, Chapter 1.
6. Johnson, Palmer O., and Robert W. B. Jackson, *Introduction to Statistical Methods*, New York, Prentice-Hall, 1953, Chapter 7.
7. Kenney, John F., and E. S. Keeping, *Mathematics of Statistics*, Third Ed., New York, D. Van Nostrand Co., 1954, Part 2, Chapters 1 and 2.
8. Mood, Alexander M., *Introduction to the Theory of Statistics*, New York, McGraw-Hill Book Co., 1950, Chapter 2.
9. Rosander, Arlyn C., *Elementary Principles of Statistics*, D. Van Nostrand Co., 1951. Chapters 6, 7, and 8.
10. Von Mises, Richard, *Probability, Statistics and Truth*, London, W. Hodge and Co. 1939.
11. Walker, Helen M., and Joseph Lev, *Statistical Inference*, New York, Henry Holt and Co., 1953, Chapter 2.
12. Wilks, Samuel S., *Elementary Statistical Analysis*, Princeton, N.J., Princeton University Press, 1949, Chapters 4 and 6.

CHAPTER 6

The Normal Distribution Function

One of the most fascinating things about the study of statistics is how practical statistical "theory" can be. There is no better example of this than the theoretical distribution which is the subject of this chapter. Though it is undoubtedly the most frequently used and the most important theoretical distribution to the statistician, it is probably, by the same token, the most frequently misused. It is variously called the normal curve, the normal distribution, the normal law, the normal probability distribution, the normal probability curve, and the normal curve of error. Some of these designations are related to the historical development of the mathematics of this distribution. It is still sometimes called the Gaussian curve, for instance, after Gauss, who was presumed initially to have developed it, and who used it as a *law of error* as did Laplace and others in astronomical observations.

It is not strange that the adjective "probability" is used in connection with the normal distribution. A mathematician, DeMoivre, originally appeared to have developed it as an outgrowth of his interests in gambling and games of chance. For historical reasons we shall use the term *normal* for this bell-shaped distribution, although, as we already know from frequency distributions which we have studied, it is quite *normal* for many types of measures not to be distributed in the normal form. For a special reason we will call it the *normal distribution function*, or simply the *normal distribution*. Our reason for this will not be thoroughly clear until in later chapters we have made use of distribution functions in the solution of statistical problems.

Before learning about the important properties of this theoretical distribution we should refresh ourselves on a concept from elementary algebra which is indispensable in the study of many topics in statistics.

6.1 THE FUNCTION CONCEPT

If we know the radius of a circle, we can compute its circumference. If we know the distance between two points on a navigation map, and if we know our ground speed, we can estimate the time it will take to fly an airplane from one point to the other. The circumference depends upon the diameter, and the time depends upon the distance. Similarly, measures with which we deal in education often may be thought of as depending one upon another. For instance, we compute the I.Q. as a ratio of mental age to chronological age. Therefore, we may say that a youngster's I.Q., at a given age, depends upon his mental age. Frequently some relationship like this is suspected although the formula defining the relationship has not been specified. For instance, it seems reasonable that the number of teachers in a school would depend upon the enrollment. As a matter of fact, empirical tabulations of the relationship between the number of teachers and the enrollment in schools have shown this to be true.[1] Of course, the number of teachers in a school might depend on many different things, some of which could be measured, and some of which might not be measured. Not only would it depend on the number of pupils but perhaps the resources of the community, the kind of educational program, and so on.

Another way of stating the relationships which we have been discussing is to say that the area of a circle is a *function* of its radius, the time of travel of an airplane is a *function* of the distance; that an I.Q. is a *function* of mental age; and that number of teachers in a school is a *function* of enrollment and other factors. The central idea is that we can determine one of two values, once we have determined the other.

Here are some functions which you might recall having studied in elementary algebra:

A linear function: $aX + b$

A quadratic function: $aX^2 + bX + c$

A cubic function: $aX^3 + bX^2 + cX + d$

An exponential function: ka^{-X}

The letters a, b, c, d, and k represent constants. Once they are specified, X values can be substituted, and the magnitudes of each of these functions of X can be determined.

By means of the conventional Cartesian coordinates we can plot a curve representing the "function of X." Thus, for an example of a linear function of X we may write:

$$Y = (5/4)X - 5$$

[1] The study by Paul R. Mort, *The Measurement of Educational Need*, Teachers College, Columbia University, 1924, was the first of a long series of studies which made use of the idea of this section in developing empirical formulas which considerably advanced the field of public-school finance.

Or we could use the symbol $f(X)$ for "function of X," and write

$$Y = f(X) = (5/4)X - 5$$

For that matter, it is not necessary that we use Y at all. Instead the symbol $f(X)$ can be used in place of Y to represent the function of X.

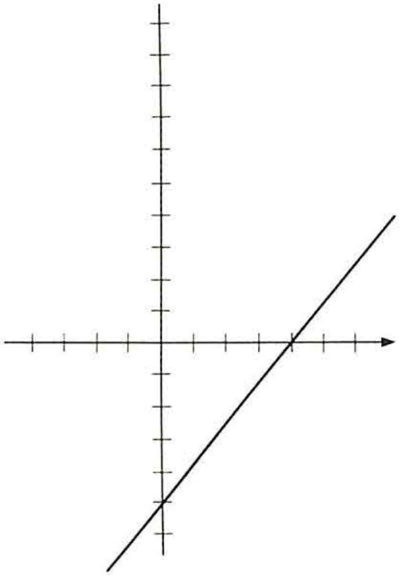


FIG. 6.1. Graph of $f(X) = (5/4)X - 5$.

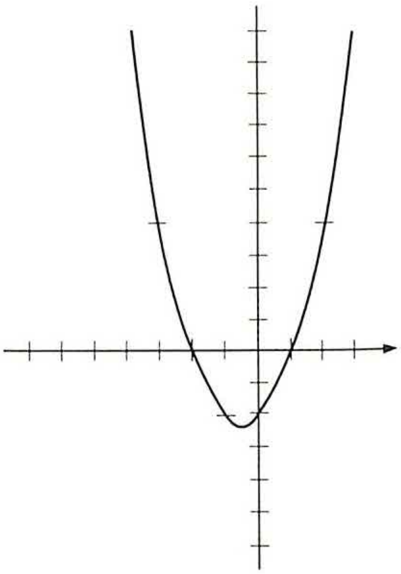


FIG. 6.2. Curve of $f(X) = X^2 + X - 2$.

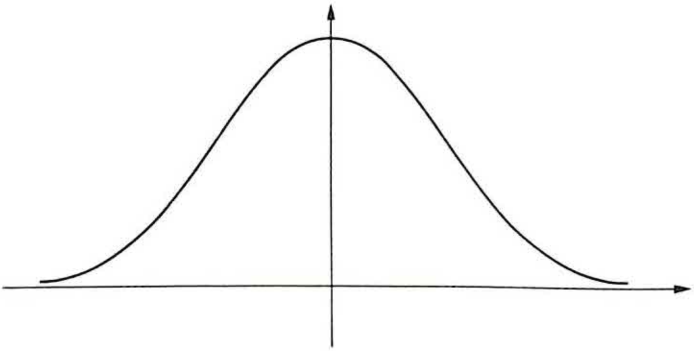


FIG. 6.3. Curve of $f(X) = e^{-X^2}$.

A *linear* function is a straight line as we see in the "curve" of this function in Fig. 6.1. The curve for the function $X^2 + X - 2$ appears in Fig. 6.2. In Fig. 6.3. is the curve of the function e^{-X^2}.

Only a slight modification of the function whose curve is shown in Fig. 6.3 will give us the function of the normal distribution. As a matter of fact, all that is necessary is a change in the scales on the chart. But no scales are given in Fig. 6.3, so that if we may use, on the X axis and on the $f(X)$ axis, whatever scales we choose, we can, in fact, take Fig. 6.3 to represent for us the normal distribution.

Changing scales is equivalent to multiplying both X and the function $f(X)$ by suitable constants. In other words, the normal distribution is of the following type:

$$f(X) = ae^{-bX^2} \tag{6.1}$$

6.2 THE NORMAL DENSITY FUNCTION

If we use some standard unit of X, independent of the magnitude of the measuring instruments used, in the place of X in the formula, we will have another form of equation (6.1). We saw in Section 4.7 that a standard unit of measure was the deviation of X from the mean divided by the standard deviation. This is the *standard score*, $z = (X - \mu)/\sigma$. Then, if for the value b we substitute 1/2 and for the value a we substitute $1/\sqrt{2\pi}$, the result is the *standard form* of the normal frequency distribution:

$$f(X) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \tag{6.2}$$

where π is a constant approximately equal to 3.1416 and e is the base of Napierian logarithms, about 2.7183. We see that the form of the function as given in equation 6.2 is in terms of X in *standard deviation units*. The coefficient of e is .3989. This is the value of a in equation 6.1. It is to be noted that the value of b in equation 6.1 is $\frac{1}{2}$ that in equation 6.2.

An important feature of the function in equation 6.2 is that, if we draw a curve of $f(X)$, *the area under it is* 1.000. Figure 6.4 is such a diagram. It is a *continuous* curve. Theoretically the points on it would be located from a great many possible values of X (or in our units z) and the corresponding ordinates, $f(X)$. In this respect, we may choose to look upon it as a probability distribution, or *probability density* as it is often called, as we did the histogram in Fig. 5.2.

Since we measure X in terms of z (deviations expressed in sigma units), all values of X above the mean are to the right of the zero point on our curve; all values less than the mean are to the left. Therefore, the reference point from which to measure z distances in Fig. 6.4 is at zero. It represents the location of the mean of the distribution, whatever that may be in the original units. In the diagram, of course, it is zero, because all values are measured in deviation units. Such a curve could be plotted by actually computing values of the function for various values of $(X - \mu)/\sigma$. Because of frequent occasion to use this function in statistical work, tables have been prepared giving "ordinates of the normal curve" for various values of z. Such a table of ordinates appears in Appendix C.

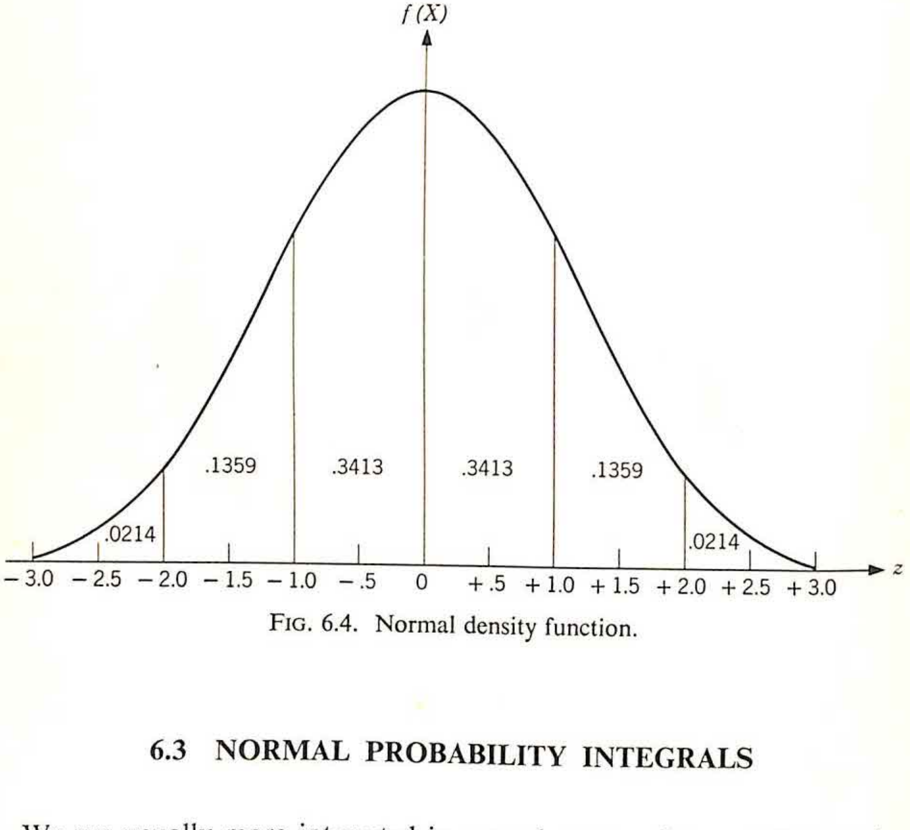


FIG. 6.4. Normal density function.

6.3 NORMAL PROBABILITY INTEGRALS

We are usually more interested in areas between given z values under this curve than we are in the ordinates. Area in probability functions and in frequency functions is generally expressed as a proportion or as relative frequency. In Tables 3.1 and 5.1 we had relative frequencies, one of an actual frequency distribution and the other of a binomial distribution. They were *discrete* distributions, and it was a simple matter to find area under parts of their respective histograms. We discovered that we could find the area between and including two X values merely by summing the relative frequencies between and including these X values. In symbolic form this may be written

$$\sum_{x=a}^{b} P(X_i) \qquad \text{or} \qquad \frac{\sum\limits_{x=a}^{b} f_i}{N} \tag{6.3}$$

This might represent finding the area between (and including) two points in the point binomial, a discrete distribution as shown in Table 5.1. From that table, the sum of the relative frequencies, or probabilities, between two heads and six heads, inclusive, is .81739. In symbols this may be written

$$P(2 \leq X \leq 6) = .81739$$

In this case $a = 2$ and $b = 6$. We are summing between (and including) the points 2 and 6. By virtue of our definition of number (see Section 2.4) the limits might be considered to be 1.5 and 6.5. In the present case only a discrete number of heads may appear, so the limits are to 2 to 6, inclusive.

In dealing with a smooth continuous curve an analogous method is used. Imagine very fine measurement which would permit the use of very small class intervals. The smooth curve is visualized as having the stair-step features of a histogram but in a very minute sense. The mathematician carries this process of making narrower and narrower classes to its ultimate conclusion in the *calculus*. This branch of mathematics is beyond the scope of this book, but we should familiarize ourselves with the notation which comes from the calculus since it is more or less standard in statistical literature.

By means of a process of the calculus called *integration* it is possible to find the area under any part of a curve. This is as if the area were made up of very narrow columns of width (ΔX), or (dX), and whose height would be the ordinates $f(X)$. The area of such a column would be nearly $f(X)\,dX$. The operation is thus similar to summing the area of frequency columns in a histogram. The symbol for doing this is the integration sign, $\int$. It means "the sum of" when dealing with *continuous* theoretical functions just as Σ means "the sum of" relative to *discrete* frequencies. Thus the area between points a and b in Fig. 6.5 would be written

$$\int_a^b f(X)\,dX \tag{6.4}$$

and is called the integral of $f(X)$ from a to b. The area shown in Fig. 6.5 is that area between the ordinates $X = a$ and $X = b$ and bounded by the curve $f(X)$ and the X axis. We did not need to say "between a and b, inclusive" because in the continuous case X may take on any value within the limits of the distribution, and the widths of intervals, (dX), are considered to be infinitesimally small.

We see that the integral sign in connection with continuous curves has the same meaning as the summation sign with discrete distributions. Recalling that the total area under the curve is equal to unity, and that any proportion of it would, therefore, represent the proportion of area, relative frequency, or probability, we may write

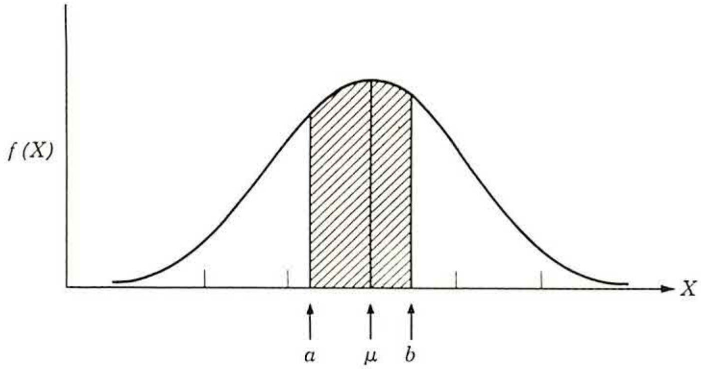


FIG. 6.5. Area under the curve.

$$P(a < X < b) = \int_a^b f(X)\,dX \tag{6.5}$$

the probability that X falls between any two values a and b. Therefore, the left-hand member of equation 6.5 may be read, "the probability that X is larger than a and less than b." This is just another way of saying that "the probability that X is between the limits a and b."

Because area is evaluated by means of an integral, and because an area of a distribution curve is a probability, these integrals are sometimes called *probability integrals*. Tables appear in most statistical books for the probability integrals of the normal curve. The table is entered according to z, the *standard deviate* from the mean. The mean, of course, as in our diagrams, is zero on the base line scale. Because the curve is symmetrical, values on the right or *positive* side equal corresponding values on the *negative* or left side. The integrals are usually given for only one side of the distribution and show areas either from the mean or zero point up to a given value of z, or areas from z up to infinity, the upper limit of the distribution. In the usual symbolism, we would find the following tabled values for the two parts of the area between a and b in Fig. 6.5:

$$P(a) < X < 0) = \int_a^0 f(X)\, dX$$

and

$$P(0 < X < b) = \int_0^b f(X)\, dX$$

Their sum is the desired integral from a to b. As we have already seen, the total area of a theoretical frequency distribution is equal to 1, or

$$\int_{-\infty}^{\infty} f(X)\, dX = 1 \qquad (6.6)$$

The symbol $-\infty$, "negative infinity," represents the left, or lower, limit of the distribution and ∞, the positive or upper limit of the distribution.

Let us suppose that $(b - \mu)/\sigma = .45$ and that $(a - \mu)/\sigma = -.75$ in the diagram of Fig. 6.5. We find in the table of Appendix C the integrals corresponding to these two z values to be respectively .17364 and .27337. If these are added together, the total proportion of area under the curve between a and b is .44701, or about 44.7 percent. Remembering that the table yields values for the function in *standard form*, we may express the results of this operation briefly in the following way:

$$\int_a^0 + \int_0^b = .17364 + .27337 = .44701$$

6.4 THE NORMAL CUMULATIVE DISTRIBUTION FUNCTION

In Fig. 3.3 and in Table 3.1 we interpreted an actual frequency distribution *cumulatively*. Similarly we may be interested in the cumulative relative frequencies (or probabilities) of a theoretical distribution. A graph of the *cumulative normal distribution function* (or simply the normal distribution function), $F(X)$, is shown in Fig. 6.6. It is comparable to the cumulative frequency or ogive curve of the 500 scores of Fig. 3.3. In the present case, however, all abscissa values are in z or sigma units so that our theoretical values may be readily converted to any of various units of measurement. The ordinates on the vertical scale, $F(X)$, show the proportion (or, if we choose, percentage, simply by shifting the decimal two places) of cases below a given z score; that is, it shows for a given z score the proportion of area under the normal *frequency* curve or normal *density* function (Fig. 6.4), also called the probability density (or probability integral), between the lower limit of the distribution and the given z score.

Since the normal (frequency) function is symmetrical, it is not necessary to give values for $z < 0$. For instance, in the table of Appendix D we find that for $z = 1.00$, that is, a score of $1\ \sigma$ above the mean, $F(X) = .8413$. This represents the proportion of the area lying *below* $z = 1.00$. The remaining area, the area *above*, is $1.0000 - .8413 = .1587$. Moreover, the area or probability *above* $z = 1.00$ is the same as the area *below* $z = -1.00$. Hence the $F(X)$ value for $-1.00z$ is .1587.

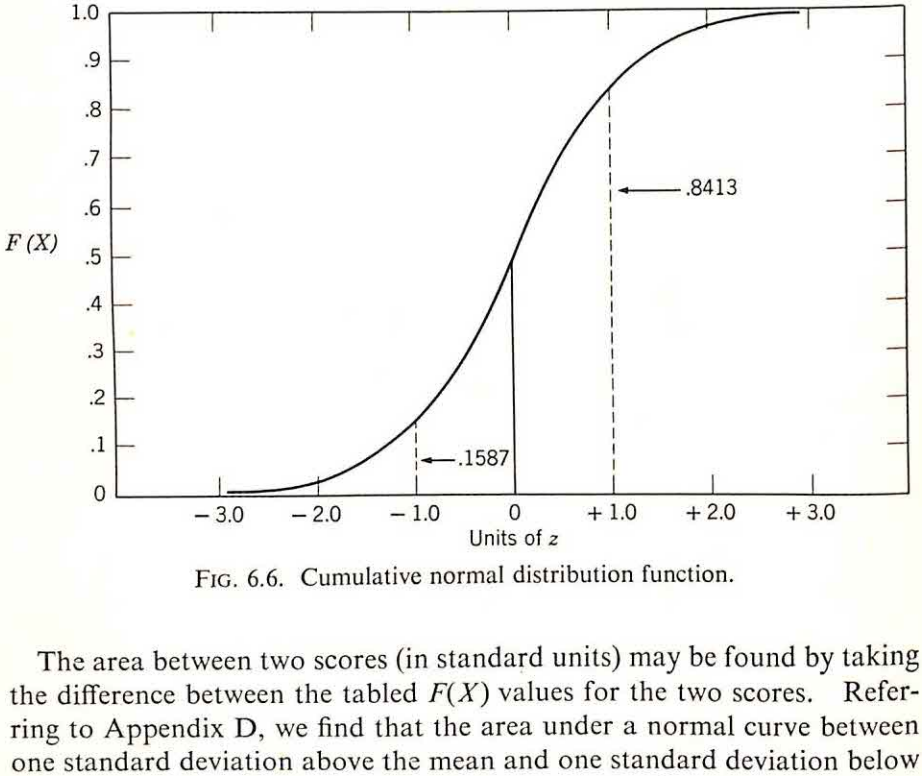


FIG. 6.6. Cumulative normal distribution function.

The area between two scores (in standard units) may be found by taking the difference between the tabled $F(X)$ values for the two scores. Referring to Appendix D, we find that the area under a normal curve between one standard deviation above the mean and one standard deviation below the mean is $.8413 - .1587 = .6826$, or about two-thirds.

We may thus find areas under the curve in two ways: (1) by means of integrals of the frequency or density function, $f(X)$, or (2) by means of the cumulative distribution function, $F(X)$.

For any specified value of X, say a, $F(a)$ is the probability that $X < a$, that is,

$$F(a) = P(X < a) \tag{6.7}$$

We noted above that differences between tabled values of $F(X)$ give areas under the *frequency* curve, or the probability that X falls between the two values. Thus

$$\begin{aligned} F(b) - F(a) &= P(X < b) - P(X < a) \\ &= P(a < X < b) \end{aligned} \tag{6.8}$$

We should be able to satisfy ourselves that

(1) $F(-\infty) = 0$

(2) $F(\infty) \quad = 1$

(3) If $a < b$, then $F(a) < F(b)$

The table in Appendix D may be used in conjunction with or in place of the table of Appendix C for finding areas (probabilities or relative frequencies) under the normal curve. As we have seen, the integrals in Appendix C are areas from the mean to a specified value of z, $\int_0^z$, whereas the integrals of Appendix D are areas from the lower limit to a value of z, $\int_{-\infty}^z$. In the normal curve the area below the mean is .50. Hence, the entry in Appendix D for a specified value of z is equal to .50 more than the integral of Appendix C for that value of z. For example, from Appendix C, $\int_0^{1.25} = .3944$ where the specified value of $z = 1.25$. In Appendix D we find the entry for $z = 1.25$ to be $F(X) = \int_{-\infty}^{1.25} = .8944$. This is .5000 more than .3944.

Either table may be used to find a specified percentile value of z. We will use a subscript to identify the percentile sought. Suppose that we wish to find the eightieth percentile, $z_{.80}$. In Appendix C, $z_{.80}$ will correspond to $\int_0^{z_{.80}} = .30$. The nearest tabled value to .30000 is .29955, the integral for $z = +.84$. In Appendix D, the z value corresponding nearest to $\int_{-\infty}^{z_{.80}} = .8000$ is also found to be $+ .84$. Hence, in that table also $z_{.80} = + .84$.

To find percentiles of z less than the fiftieth, it is necessary in either table to proceed as if the entries apply to the *opposite side* of the distribution. For instance, from Appendix C we may find $z_{.20} = -z_{.80}$

$= -.84$ since $\int_{z_{.20}}^{0} = \int_{0}^{z_{.80}} = .30$. From Appendix D also we find $z_{.20}$ as the negative of the value of $z_{.80}$ since

$$\int_{-\infty}^{z_{.20}} = \left(1 - \int_{-\infty}^{z_{.80}}\right) = \int_{z_{.80}}^{\infty} = .20$$

In learning these relationships it is helpful to sketch a normal density function as in Fig. 6.4, marking off values of z on the abscissa and locating z values sought and identifying related areas or relative frequencies.

6.5 THE NORMAL APPROXIMATION TO THE BINOMIAL

In connection with our hypothetical teacher prediction experiment in Chapter 5, we made some probability statements based on the binomial. In Table 5.1 we reported the probabilities for the binomial with $N = 10$ and $p = 1/2$. We added the probabilities for $X = 8$, $X = 9$, and $X = 10$ to find the chances of getting 8 *or better*. This was found to be .0547. We had decided in advance that the "null hypothesis" would be rejected if the probability of the observed outcome would be .05 or less, $P \leq .05$.

We shall now see that the normal distribution would have given us a reasonable approximation to the binomial probability. In order to "fit" the normal to the binomial distribution, we must first compute the mean and standard deviation of the binomial distribution. These we found in Section 5.5. From equations 5.8 and 5.9, they are

$$\mu = Np = 5.0$$

$$\sigma = \sqrt{Npq} = 1.58$$

Next we must specify that area under the normal curve which corresponds to the three rectangles for the values of 8, 9, and 10 in the histogram of our binomial (Fig. 5.2). Remembering that the normal distribution is *continuous*, we see that values of 8 *or more* in the binomial will correspond to values of 7.5 or more in the normal distribution. Since tables of the normal distribution are entered in *standard* units, we now find the standard score equivalent to 7.5:

$$z = (X - \mu)/\sigma = (7.5 - 5.0)/1.58 = 1.58$$

Therefore, our *normal probability* will be the integral of the normal function, from 1.58 to ∞, which is the area of the right-hand tail beyond, that is, to the right of, $z = +1.58$. From Appendix D we may find the area to the left of, that is, below $+1.58$, $\int_{-\infty}^{1.58} = .9429$. Hence,

$$\int_{1.58}^{\infty} = P(X > 7.5) = 1.0000 - .9429 = .0571$$

The result is not seriously different from the exact probability of .0547 from the binomial expansion.

From this it appears that the normal curve is not a bad approximation to the binomial. As a matter of fact, it may be shown that, even if p and q are unequal, the limiting form of the binomial, as N becomes large, is the normal function. Caution must be exercised where N is small, say less than 30, if p and q are unequal, and particularly if probabilities near the tails of the distribution are to be approximated. Nevertheless, where

TABLE 6.1

COMPARISON OF BINOMIAL AND NORMAL PROBABILITIES—BINOMIAL DISTRIBUTION $p = .3$ and $N = 20$

X	Probability, $P(X)$	
	Exact Binomial	Normal Approximation
15	.00004	.0000
14	.00022	.0001
13	.00102	.0006
12	.00386	.0029
11	.01201	.0104
10	.03082	.0298
9	.06537	.0674
8	.11440	.1209
7	.16426	.1715
6	.19164	.1928
5	.17886	.1715
4	.13042	.1209
3	.07160	.0674
2	.02785	.0298
1	.00684	.0104
0	.00080	.0036
Total	1.00001	1.0000

$\mu = 6.0; \quad \sigma = 2.049$

$p = q = .5$, as in the binomial of Section 5.3 with $N = 3$, the exact and the normal probabilities are close:

X	Probabilities, $P(X)$	
	Actual	Normal
3	.125	.124
2	.375	.376
1	.375	.376
0	.125	.124

However, returning to the previous example where $N = 10$ and $p = .5$, we find that the normal approximation, $P(X > 9.5)$ is .00221 as compared with the exact probability from the binomial, $P(10) = .00098$.

A visual inspection of the histogram of Fig. 5.3 suggests the *goodness of fit* of the normal curve to the binomial distribution with $p = .3$, and $N = 20$. The probabilities may be compared in Table 6.1.

6.6 FITTING THE NORMAL CURVE— THE AREA METHOD

One reason the normal distribution is important in statistical work is that many formulas and statistical devices are neatly derived by assuming normality. As we have seen, for the determination of either areas (probabilities) or ordinates, all we need to know about a normal distribution is the mean and the variance, or standard deviation. Then, of course, if actual frequencies are desired, only the additional value, N, needs to be known. These three values which determine any given normal distribution, N, μ, and σ, are *constants* for a particular distribution. Other distributions have their own appropriate constants. We have also seen that the normal distribution is the limiting form of the binomial. Recalling that the binomial represented a distribution of random events, we see that the normal distribution offers a description of the operation of *chance*. We shall see later how further concepts of statistical description and sampling will use the normal distribution.

It is often important to compare an actual frequency distribution with the normal by "fitting" the normal curve to it. If the fit is good, the application of a technique which assumes normality of distribution is justified. One method of doing this is to compare the actual frequency in each class interval of an observed frequency distribution with those which would result from the best-fitting normal curve. In fitting a normal curve to a frequency distribution, (1) the mean of the theoretical curve is set equal to the mean of the frequency distribution, (2) the area of the curve is made equal to the area of the histogram, and (3) the standard deviation of the curve is made equal to the standard deviation of the frequency distribution.

After the normal distribution is fitted, comparison with the observed distribution requires that the integrals *between the limits* of each class interval be evaluated. They are the proportions of area or total frequency, N, expected within each class.

TABLE 6.2

COMPARISON OF OBSERVED AND NORMAL FREQUENCIES, SCORES OF 500 NAVAL RECRUITS

	Frequencies	
Score	Observed	Normal
95–99	—	0.7
90–94	3	1.4
85–89	4	3.4
80–84	11	7.4
75–79	13	14.4
70–74	18	24.9
65–69	37	38.4
60–64	50	52.4
55–59	72	63.6
50–54	70	68.6
45–49	56	65.5
40–44	58	55.9
35–39	50	42.2
30–34	30	28.3
25–29	13	16.9
20–24	9	9.0
15–19	4	4.2
10–14	2	1.8
5–9	—	1.0
Total	500	500.0

This was done in our fitting of the normal curve to the binomial when we converted binomial values to standard form such that $\mu = 0$ and $\sigma = 1$, the same as tabled z values for the normal distribution.

To illustrate the procedure further, we will use the distribution of 500 scores of Table 3.1. The interval, 60–64, was reported to have a frequency of 50. The limits of this interval are 59.5 and 64.5. The mean and standard deviation are $\bar{X} = 51.33$ and $\sigma = 14.46$. The standard deviation units of the limits are $(59.5 - 51.33)/14.46$ and $(64.5 - 51.33)/14.46$, or $z' = .565$ and $z'' = .911$. By interpolation in Appendix D we find the respective cumulative functions to be $F(X') = .7140$, and $F(X'') = .8189$. From equation 6.8 we find that

$$P(59.5 < X < 64.5) = .8189 - .7140 = .1049$$

the *relative frequency of cases in a normal distribution with mean of* 51.33 *and standard deviation of* 14.46 *between the values* 59.5 *and* 64.5. Since we are interested in a distribution with $N = 500$, we must multiply the relative theoretical frequency by 500 to get the theoretical frequency of 52.4 This we may compare with the *observed* frequency of 50. Repeating this operation for each of the intervals of the distribution yields the results shown in Table 6.2.

6.7 GRAPHING THE FITTED NORMAL CURVE

To graph the fitted normal curve on the histogram of an observed frequency distribution, it is convenient to modify equation 6.2 for the normal frequency function. To this point in our discussion of the normal distribution we have considered it only as a distribution of *unit variance* and *zero mean*, that is, in terms of standard units. In the computations in Table 6.2 it was necessary to translate the relative frequencies into actual frequencies by multiplying by 500. In modifying equation 6.2 to give ordinates of a curve whose area is N instead of 1.00, we must have N times the value given by equation 6.2.

In addition, the ordinate of the curve depends upon the dispersion—how much the curve is spread out. In fact, the greater the standard deviation, the smaller the ordinate. This notion should be evident from a scrutiny of Fig. 6.7. Here we have two normal curves of equal area, $N = 1{,}000$. Curve A has a standard deviation of 10; curve B, a standard deviation of 20. Clearly the area of curve B is *spread out* twice as much as the area of curve A. Therefore, the ordinates of curve B at any given value of z should be half the ordinate of curve A. The arrows point to the respective ordinates at $z = 1$ although, to avoid confusion, they are not drawn in the chart. They may be seen to be in the ratio of 2 to 1. At $z = 0$, that is, at the mean, it may be seen that the "maximum" ordinate of A is twice that of B.

The adjustment for N and σ may be combined into one factor, (N/σ). Thus the ordinate for a "fitted" curve which is to correspond to a given N and σ will vary *directly* as N, and *inversely* as σ.

In Fig. 6.7 a visual check may be made of the correctness of multiplying by (N/σ). The ordinate at $z = 0$ for the frequency function (for which $N = 1.00$) is .3989, from Appendix C. The ordinate for curve A is $.3989(1{,}000/10) = 39.9$; for curve B is $.3989(1{,}000/20) = 19.9$. By this process ordinates under the arrows ($z = +1.0$) may be found to be 24.2 and 12.1 respectively. By evaluating a number of ordinates, a smooth curve of the fitted distributions can be graphed. Finally, if the diagram has been made on coordinate paper, a count of squares under each curve will show that both areas are 1,000.

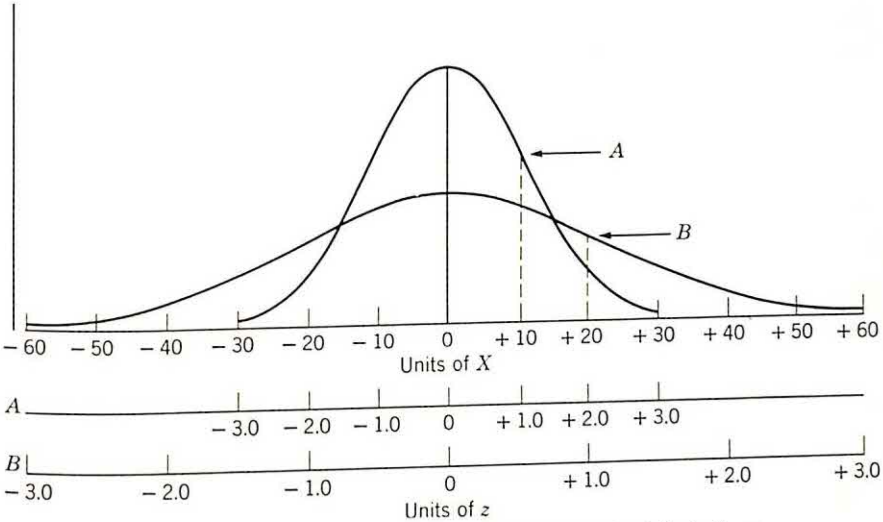


FIG. 6.7. Curves of equal area—different standard deviations.

Usually there is still a further consideration in finding ordinates for fitted frequency distributions. More often than not the frequencies are tabulated in some class interval u other than one. A histogram in raw scores will have frequencies (that is, ordinates and hence areas) five times as great if $u = 5$ than if $u = 1$.

This results from making a block 5 units wide and 50 units high to represent 50 observations, when really there are only about 10 observations per unit of raw score. Therefore, we must also multiply further by u in order to convert equation 6.2 into units comparable to those in which the observed frequencies appear in the histogram. Making all of these adjustments, we can write the formula for the ordinate of the normal curve, *adjusted* to the scale of an observed frequency distribution,

$$Y = \left(\frac{uN}{\sigma}\right) f(X) \tag{6.9}$$

Let us see how this would be applied to the distribution of Table 6.2. The histogram for this frequency distribution was shown in Fig. 3.1. In this distribution $N = 500$, $u = 5$, and $\sigma = 14.46$. In this case, we would find from equation 6.9 that

$$Y = [(5)(500)/14.46]f(X) = 172.9 f(X).$$

We find $f(X)$ for the *unit* normal curve in Appendix C for corresponding values of z. At the mean, which in this case is 51.33, $z = 0$. From the table of ordinates we find for $z = 0$ that $f(X)$ is .3989. Therefore, the ordinate at the mean of the normal curve which best fits our histogram is $172.9 \times .3989 = 69.0$. In the histogram, $z = 1$ corresponds to $51.33 + 14.46 = 65.8$. From the table of ordinates we find that at $z = 1$ the ordinate is .2420. Multiplied by the adjustment factor, 172.9, this gives us 41.8, the ordinate at $\mu + 1.00\sigma$ of the normal curve adjusted to fit our distribution. Since the curve is symmetrical, the adjusted ordinate for $51.33 - 14.46 = 36.9$ is also 41.8.

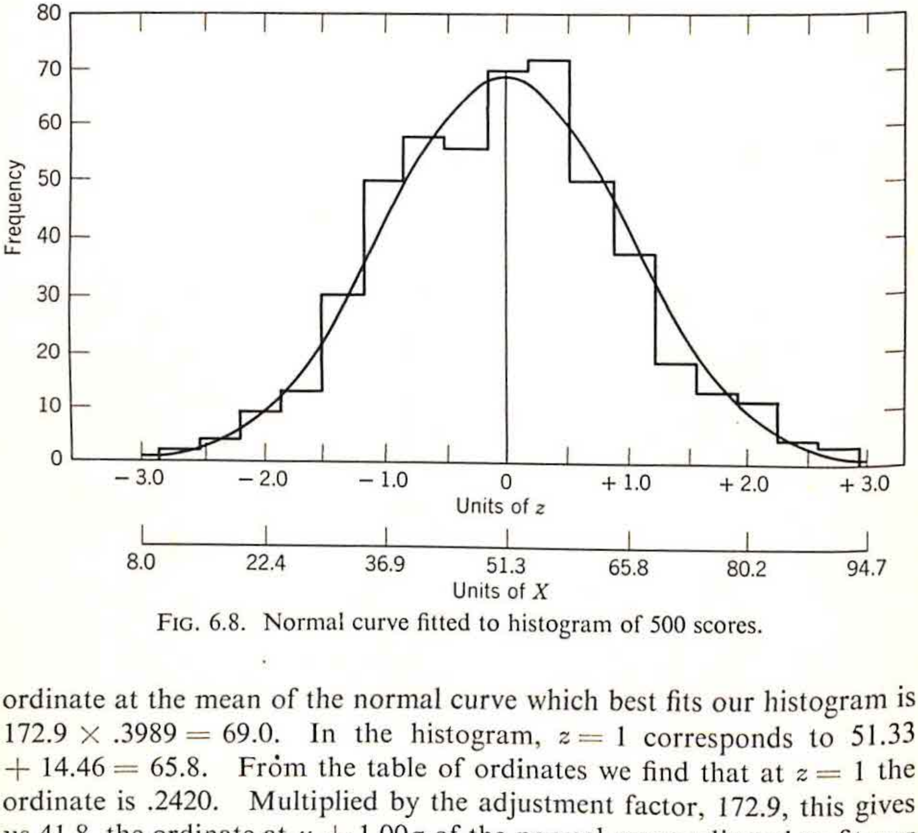


FIG. 6.8. Normal curve fitted to histogram of 500 scores.

By repeating this process for several z values a sufficient number of points may be located on the histogram to permit sketching the normal curve. Several such points for the distribution of 500 scores are shown in Table 6.3. The histogram and the normal curve are shown in Fig. 6.8.

TABLE 6.3

SOME OF THE ORDINATES OF THE NORMAL CURVE IN FIG. 6.8*

Standard Score, z	Raw Score, X	Normal Function, $f(X)$	Adjusted Ordinate, Y
(1)	(2)	(3)	(4)
+3.0	94.7	.0044	0.8
+2.5	87.5	.0175	3.0
+2.0	80.2	.0540	9.3
+1.5	73.0	.1295	22.4
+1.0	65.8	.2420	41.8
+ .5	58.6	.3521	60.9
0.0	51.3	.3989	69.0
− .5	44.1	.3521	60.9
−1.0	36.9	.2420	41.8
−1.5	29.6	.1295	22.4
−2.0	22.4	.0540	9.3
−2.5	15.2	.0175	3.0
−3.0	8.0	.0044	0.8

* Column 1: Arbitrarily chosen.
Column 2: 51.33 + 14.46 (Col. 1).
Column 3: Ordinates from App. C. corresponding to z values in Col. 1.
Column 4: 172.89 (Col. 3).

6.8 THE NORMAL CURVE AND EDUCATIONAL MEASUREMENT

In subsequent chapters we will encounter numerous ways in which the normal distribution enters into the development of statistical methods applicable to the study of educational problems. In Chapter 7, for instance, we shall make use of the normal distribution in learning about some of the most strategic statistical ideas connected with problems of sampling. For the present we will discuss some of the applications of the normal distribution which have become quite common in educational measurements.

It is important to remember in all of these applications that the normal distribution is a *mathematical model*, an idealized or theoretical description of a universe of measures. This should be understood by anyone who has followed the major theory of this chapter to this point.

The normal distribution has often been improperly used. Unfortunately there have been those who have tended to make a fetish of it, a panacea, a standard or criterion to which everything should conform. Nevertheless it is one of the devices which has permitted phenomenal progress in the field of educational and psychological measurement, and the examples in the remainder of this chapter should be viewed in that light.

The purpose of the following sections is to explain some of the more common techniques used in test construction, scaling, and educational practice which are based upon normal distribution theory. The proper application of these techniques depends upon such considerations as the educational propriety of the particular plan of action which may be intended, the validity and reliability of the measures which are to be used, considerations of sampling theory which may be inherent in the problem, and finally, the reasonableness of the assumption that the trait or attribute under consideration is normally distributed. These are in great part matters of educational theory and practice or of measurement theory and practice beyond the scope of this book.[1]

6.9 DIVIDING A GROUP OF MEASURES INTO A GIVEN NUMBER OF SUB-GROUPS OF EQUAL RANGE OF TALENT

Suppose that an attribute such as general ability, mechanical aptitude, proficiency in reading, or some other characteristic of a population is *assumed to be normally distributed.* On the basis of some measurement, that is, rankings by a teacher, test scores, it is possible to arrange individuals in the population in order. If it is desired to distribute the individuals into six groups of "equal range of talent" in the attribute, the expected proportions will be as in Fig. 6.4, each class being assigned a σ interval (z units) on the base line.

Class	A	B	C	D	E	F
Percent of Total	2.1	13.6	34.1	34.1	13.6	2.1

The range -3σ to $+3\sigma$ includes almost all members of the population. We could divide it into *five* equal distances of 1.2σ units each. The five

[1] For a more extended discussion of applications of the normal distribution to educational measurement see references 3, 4, and 7. An understanding of the materials in this chapter should greatly facilitate the reading and comprehension of such sources.

classes would be determined by the limiting σ or z values in the following table. The percentages are from the table of normal integrals.[1]

Class	Limits on σ Units	Percent of Total
A	-3.0 to -1.8	3.5
B	-1.8 to -0.6	23.8
C	-0.6 to $+0.6$	45.2
D	$+0.6$ to $+1.8$	23.8
E	$+1.8$ to $+3.0$	3.5

Since in the normal distribution almost 99 percent of cases are between $\pm 2.5\sigma$, sometimes that range is divided into five classes, with one standard deviation of range each.

The foregoing example will be recognized as the familiar "normal curve" method of assigning marks, a practice which may be questioned on several counts. However, modifications of this procedure have proved useful in large-scale testing and classification programs. One used by the U. S. Air Force, called the *stanine scale* (standard nine), is a system of reporting converted test scores into nine normalized categories, 1 to 9. The mean of this system of scores is 5, and each scale value has a range of $\sigma/2$.

Several other systems of transformed scores are in use. Some are *linear transformations* of scores, not *area transformations* such as those already discussed. The commonest type of linear transformation is equation 4.16, discussed in Section 4.7 in connection with standard scores. It is simply a conversion of raw scores to a derived scale, with a mean of 50 and standard deviation of 10. Such transformations do not change the shape of the original distribution. Measures of skewness and kurtosis from these derived scores will be equal to those of the raw scores. In a positively skewed distribution, it would thus be quite possible for Z scores to be as high as 90 or 100, for instance, whereas in the normal distribution rarely would scores be expected above 80. By the same token, such a distribution might have very few scores below, say, 35. In order that norms or derived scores on tests have comparable meanings

[1] People are more accustomed to thinking in terms of percents than of proportions. The terms *percent* and *percentage* will be used frequently in this book when in fact reference is made to proportion. In statistical work it is convenient to be able to think "23.8 percent" even though the table is .238. Since *percentile* may be defined as points below which there are $100 f/N$ *percent* of the cases, they may also be considered as simply points cutting off the lower f/N *proportion* of cases. We will therefore use *percent* and *proportion* as synonyms, realizing that it is necessary to be careful in "multiplying by 100, or setting the decimal over two places" when we *really mean percent*. It is a good practice to keep all data on work sheets in terms of proportion to avoid errors caused by misplaced decimals.

the area method of transformation is frequently used. It involves "normalization" of the distribution. Regardless of the shape of the original distribution, the derived scores will be normal.

6.10 NORMALIZING SCORES

Of the several types of schemes for normalizing scores, the allocation of ordered individuals in a population according to area of the normal curve is the simplest. This may be done from tables of integrals of the normal curve above, or it may be done graphically. Only a slight modification of the chart of Fig. 4.4 would be necessary to make it a conversion chart yielding normalized scores for the General Classification Test used with the 500 naval recruits. All that would be necessary would be to draw in an appropriate derived scale on the vertical axis, using whatever value was to represent the mean and whatever interval was to represent one standard deviation, and marking off, above and below the mean, scale values corresponding to the appropriate cumulative percents. It is to be remembered that the chart is on normal probability paper, and the graph of a normal distribution is a straight line on it. Smoothing the plotted points on such a chart results in normalizing.

One of the earliest systems of normalizing was McCall's "*T* score." It involves first converting scores to percentiles, then *z* equivalents from a table of normal integrals, and finally from these a transformation such that the mean is 50 and the standard deviation 10. We shall illustrate the usual arithmetical method by which this is done. It is essentially equivalent to the graphical method described above.

In standardizing tests, it should be noted first of all that a fairly large number of cases, that is, a sample of adequate size, be used as the standardizing group. Otherwise much of the work in establishing norms, whether by some system of standard score transformation or otherwise, may be meaningless. Therefore, before proceeding with our description of how to compute *T* scores, we shall illustrate the magnitude of the sampling problem.

Bear in mind that the *Z* scores and the *T* scores which we are discussing have units of $.1\sigma$. We have learned enough about the binomial distribution and sampling to consider roughly how stable from sample to sample the median or fiftieth percentile, P_{50}, would be. Imagine a number of samples of size $N = 100$ from a continuous population of possible test scores. The probability of a score less than the median is 1/2, by definition. Hence the distribution of the 100 items in any sample into two groups, those above the median and those below it, is a binomial distribution with $N = 100$ and $p = 1/2$. We could compute the exact probabilities for this distribution, but we have learned that it is approximated by the normal curve. We will use this knowledge, and therefore compute the mean and standard deviation. The mean is $\mu = Np = (100)(1/2) = 50$, and

$$\sigma = \sqrt{Npq} = \sqrt{(100)(1/2)(1/2)} = 5$$

We have seen that

$$P\,[(\mu - 1.0\sigma) < X < (\mu + 1.0\sigma)] = .68$$

Therefore, whatever raw score on our test is the *true* median will occur in about two-thirds of our samples as some percentile between P_{45} and P_{55}. Also, the probability would be .32 that it would be outside this range of percentiles. It follows that a percentile at or near the median of a sample of 100 may easily be as much as 5 percentiles away from the population or "true" value.

A glance at the table of Appendix D shows that near the center of a distribution (near $z = 0$), a difference of as much as 5 percent represents more than $.1\sigma$, the unit used in the T score. It is evident, then, that satisfactory test standardization of the type under discussion may require very large samples. Furthermore, there is "error of measurement" to be taken into account. It will be treated in a later chapter, in connection with the study of correlation.

With the foregoing in mind, we should be in no danger of acquiring incorrect notions about the procedure if we use an example, for the sake of simplicity, with very low frequencies. The distribution of raw scores in Table 6.4 may be seen to depart considerably from normal. To compute percentiles for each class interval we must first find the number of cases below the midscore of the interval. This is F_m, the cumulative frequency to the midscore, that is, the frequency of one-half the interval and the sum of the frequencies of all intervals below it. For instance, the interval whose midscore is 32 has a frequency of 6. According to our definition of units of measurement, this interval includes values between 31.5 and 32.5. Half the six scores in the interval are assumed to be above the midscore, 32, and half below. There are 18 scores in the five intervals below this interval. We find the cumulative frequency to the midscore to be $18 + 3 = 21$ as shown in column 3. Dividing column 3 by $N = 46$ yields the cumulative proportions in column 4. We find the standard score, z, from Appendix D corresponding to each observed cumulative proportion. In column 6 the T score is found by computing $50 + 10z$.

A similar procedure may be followed in transforming ranks to normalized scores. In this case each rank, R, is considered a raw score for

TABLE 6.4

EXAMPLE OF COMPUTATIONS OF *T* SCORES

Score X_i	Frequency f	Cumulative Frequency to Midscore, F_m	Cumulative Proportion to Midscore*	z	*T* score
(1)	(2)	(3)	(4)	(5)	(6)
35	3	44.5	.967	+1.84	68
34	8	39.0	.848	+1.03	60
33	11	29.5	.641	+0.36	54
32	6	21.0	.457	−0.11	49
31	5	15.5	.337	−0.42	46
30	7	9.5	.207	−0.82	42
29	3	4.5	.098	−1.29	37
28	1	2.5	.054	−1.61	34
27	2	1.0	.022	−2.01	30

* The percentile value of X_i times 1/100, that is, the percentile value of 35, is 96.7; of 34, 84.8; etc.

which all the frequencies are one except where there are ties. The procedure of Table 6.4 may be followed, but it should be noted that the largest *R* represents the lowest rank. The proportion below a given rank may be computed as in column 4 of Table 6.4 or by formula,

$$P_R = 1.00 - \frac{R - .5}{N} \qquad (6.10)$$

where *R* is the rank of the individual and *N* is the number of cases. As before, *z* scores may be found for the corresponding percentiles (or proportions) from normal tables. In turn the *z* scores, which are normal deviates, may be transformed linearly to *T* scores or similar normalized scores of suitable units which avoid the inconvenience of the positive and negative signs.

The kind of thinking which calls for such a procedure assumes that if we had the proper measure, rather than the one that we really have, it would be distributed in the normal form. We may be interested in studying teachers with respect to some characteristic, let us say fair-mindedness. A number of teachers are ranked, that is, assigned numbers in order from 1 to *N*. This in itself gives us values from 1 to *N* which have little appeal because we believe that the proper dimension, if it could be measured, would be normally distributed.

One danger, of course, is in the specificity with which the dimension is defined. Rankings based upon judgment will be less reliable if there are several dimensions. Instead of a ranking of teachers on a rather general basis, more reliable results may usually be expected if ranking is made on several narrowly defined "unitary traits." Otherwise the situation is similar to that of attempting, without further specification, to rank the principal cities of the United States according to geographic location. On the other hand, if the ranking is to be done on average annual rainfall, latitude, longitude, distance from Washington, D. C., or some other specific single factor, the task is definite and straightforward.

6.11 ASSIGNING DIFFICULTY VALUES TO TEST ITEMS

Many of the techniques most used for the evaluation of test items make some use of the normal distribution. The simplest information about a test item is the relative number, the percent or proportion, passing the item. On the basis of this information we could rank items from easiest to most difficult, using this information in the arrangement of the final test. However, if what is desired is some relative measure of *how much* of the thing measured is represented by the item, neither the rank nor the percent passing is directly useful. The latter, of course, is the area to the right of a line dividing the distribution between those possessing enough of what is being measured to pass it from those who do not. It is the division of the base line into definite segments that gives us information about the function X which presumably the entire test and each individual item measures.

The simplest system of finding difficulty values of items is reading from normal tables the z score corresponding to a given proportion passing, p, such that $p = \int_{z_p}^{\infty}$.

The following are such values for five test items:

Item	Percent Passing	Difficulty Value, z
1	23.2	+0.73
5	72.8	−0.61
16	9.3	+1.32
21	61.3	−0.29
32	39.8	+0.26

6.12 THE MEAN DEVIATION OF A PORTION OF THE NORMAL DISTRIBUTION

In the computations of Table 6.4 we located z scores for the *medians* for each of the nine portions of the observed distribution. It will be recalled that we computed cumulative proportions up to the point above which and below which were half the scores in the interval. For reasons which should be obvious to the reader, different results would be derived were we to find the *mean deviation* for each portion of the distribution.

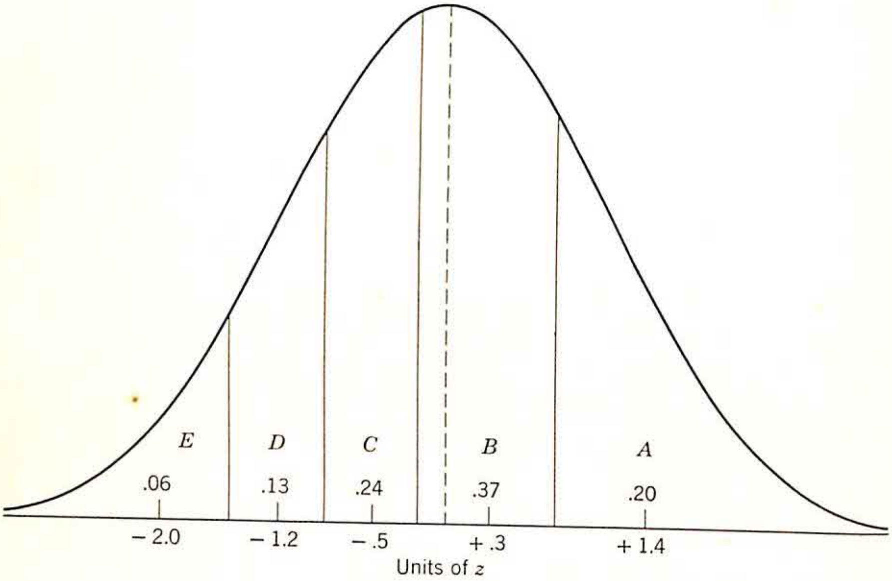


FIG. 6.9. Mean deviations from proportions in five categories.

By the calculus it can be shown that the mean deviation of a portion of the normal curve (in z units) is the ratio of the difference between the two ordinates which bound the segment and the area between them. In the notation we have been using, we may write a formula for this in several ways. One of them is:

$$\bar{z} = \frac{f(X') - f(X'')}{F(X'') - F(X')} \tag{6.11}$$

where $\bar{z}$ is the mean z score; X' and X'' are values of X which bound the interval; and the numerator and denominator are, respectively, the difference between the ordinates and the difference of the cumulative frequency functions for the two values of X. The denominator is the same as $\int_{X'}^{X''} f(X)\,dX$. We could substitute this integral in the denominator and have another way of writing the formula. Or we might let the ordinates be y' and y'' and the areas to the right of the two values be q' and q'', respectively. Then our formula would be

$$\bar{z} = (y' - y'')/(q' - q'') \tag{6.12}$$

An application of this theorem, using equation 6.11, is shown in Table 6.5 and Fig. 6.9.

The categories A, B C, D, and E shown in the first column of Table 6.5 represent some *ordered* classification. Category A is the highest, and category E is the lowest on some dimension on which there is assumed to be a normal distribution of a continuous variable which we cannot measure directly. The categories might be grades in freshman English classes, achievement or proficiency in English, or some other such variable which the grades are intended to measure. They could also be ordered categories of response to a questionnaire or opinion poll item such as:

All children should have an opportunity of attending kindergarten. (Check one)

...... A. strongly agree
...... B. agree
...... C. uncertain
...... D. disagree
...... E. strongly disagree

The proportions in column 2 of Table 6.5 are observed relative frequencies in some population, or, more likely, a sample of a population, for example, last year's freshman English class marks, or a sample of parents who have been interviewed in a school district.

In order to locate the ordinates forming the upper and the lower limits of the five areas of the curve, as shown in Fig. 6.9, we compute for each category the cumulative proportions, $F(X')$ in column 3, and $F(X'')$ in column 6. For these we look up tabled values of z as shown in columns 4 and 7. From the z scores we in turn enter the table of ordinates to find the ordinates in columns 5 and 8. We are now ready to subtract the entries in column 8 from those in column 5 to get the numerator of equation 6.11. This we enter in column 9, being careful to enter the proper sign. Finally we divide column 9 entries by those in column 2 to get $\bar{z}$ as reported in the last column, again being careful to indicate the proper sign. The resulting mean deviations for the five categories are located on the base line scale of Fig. 6.9.

TABLE 6.5

Computation of Mean Deviation of each of Five Ordered Categories, Given Relative Frequencies

Category	Proportion of Total $[F(X'')-F(X')]$	Lower Limit			Upper Limit			$f(X')-f(X'')$	$\bar{z}$
		Cumulative Proportion, $F(X')$	z	Ordinate, $f(X')$	Cumulative Proportion, $F(X'')$	z	Ordinate, $f(X'')$		
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
A	.20	.80	+ .84	.280	1.00	+ ∞	.000	+.280	+1.4
B	.37	.43	− .18	.393	.80	+ .84	.280	+.113	+ .3
C	.24	.19	− .88	.271	.43	− .18	.393	−.122	− .5
D	.13	.06	−1.55	.120	.19	− .88	.271	−.151	−1.2
E	.06	.00	− ∞	.000	.06	−1.55	.120	−.120	−2.0
Total	1.00	—	—	—	—	—	—	—	—

The preciseness of this technique is seldom justified because of assumptions of normality and the instability of proportions due to sampling. The mean deviation has been employed as a system of scoring items in a questionnaire or opinion poll. Questionnaires scored in this manner, that is, by means of $\bar{z}$ used as an item *weight*, are usually not found more suitable (for example, more reliable or valid) than an arbitrary system of weighting, such as assigning to the above categories the values 5, 4, 3, 2, and 1 as in a "point grading system" or as in most opinion measures of the type illustrated.

EXERCISES

1. Define:

Normal distribution
Function
Linear
Quadratic
Probability density
Probability distribution
Probability integral
Continuous function
Discrete function
Distribution function
T score
z score
Normalized score
Transformation of scores
Mean deviation

2. Sketch the function $f(X)$ as in Fig. 6.5 and identify areas specified in the following. Where necessary verify values given by reference to Appendix D or Appendix C.

(*a*) $\int_0^{z_{.70}} = \int_{z_{.30}}^{0} = .20$

(*b*) $P(z_0 < z < z_{.70}) = .20$

(*c*) $P(z_{.40} < z < z_{.70}) = .30$

(*d*) $P(z_{.40} < z < z_{.50}) + P(z_{.50} < z < z_{.70}) = .30$

(*e*) $\int_{z_1}^{z_2} = \int_{-\infty}^{z_2} - \int_{-\infty}^{z_1}$

(*f*) $\int_{z_{.30}}^{0} + \int_0^{z_{.80}} = .50$

(*g*) $P(z > 1.00) = P(z < -1.00) = .1587$

(*h*) $P(z > 2.00 \text{ or } z < -2.00) = .046$

(*i*) $P(z_{.15} < z_{.85}) = 1.00$

(*j*) $z_{.95} = 1.645;\ z_{.05} = -1.645$

(*k*) $z_{.975} = 1.960;\ z_{.025} = -1.960$

(*l*) $z_{.99} = 2.326;\ z_{.01} = -2.326$

(*m*) $z_{.995} = 2.576;\ z_{.005} = -2.576$

(*n*) $\int_{-\infty}^{-1.960} + \int_{1.960}^{\infty} = .05$

(*o*) $P(-.84 < z < 1.03) = .6480$

(*p*) $P(-.32 < z < .58) = F(.32) + F(.58) - 1.000 = .3445$

3. Assume that the California test scores in Appendix A are from a normal population for which $\mu = 70.00$ and $\sigma = 13.00$.

(*a*) Find the following:

$$\text{(i) } \int_{60}^{\infty}, \text{ (ii) } P(40 < X < 50), \text{ (iii) } \int_{60}^{75}, \text{ (iv) } \int_{-\infty}^{80}$$

(*b*) What score would a student have to make in a second test for which $\mu = 45$ and $\sigma = 10$ comparable to a level of maturity of a score of 90 on the California test?

(*c*) Find $P(X \leq 65)$ and $P(X > 65)$.

(*d*) Find the two scores between which the middle 40 percent of all scores would be expected.

(*e*) What is the eighty-second percentile? P_{82}? P_{24}? P_{84}? P_{16}?

4. What is the simplest method of comparing a frequency distribution with the normal distribution? (See Section 4.8.)

5. On the basis of one form of intelligence test administered to a sample of subjects, it was found that 1 percent of I.Q.'s were below 70. The standard deviation of this distribution of I.Q.'s was about 13. A later version of the test produced I.Q.'s in another sample distribution with standard deviation of 16. If *feeblemindedness* is defined as I.Q. below 70, what will be the effect of introducing the new test on the selection of handicapped children? What are some of the possible explanations for differences in the distributions of I.Q.?

6. The proportion presumed to be in favor of a bond issue in a Midwestern city is $p = .65$. By means of the normal approximation to the binomial distribution, find $P(55 < X < 75)$, where X is the number responding in favor of the bond issue in a sample poll of 100 voters.

7. From the variance Npq of a binomial frequency, prove that the variance of a proportion is

$$\sigma_P^2 = pq/N$$

8. Compute frequencies of the best-fitting normal curve for the California test scores of Ex. 1, Chapter 3.

9. Draw the normal frequency curve and the normal cumulative frequency which best fit the histogram and cumulative frequency polygon of Ex. 2, Chapter 3. Plot the cumulative frequency distribution of the California test scores on normal probability paper.

10. The proportions of the grade of A in five different courses in a university are (*a*) .08, (*b*) .17, (*c*) .24, (*d*) .15, (*e*) .27. By means of the mean deviation, find z scores which would give "scaled" values of the grade of A for each of the five courses.

11. What is the difference between the Z score from formula 4.16 and the T score of Section 6.10?

12. Given two distributions on an achievement test, one for high-school freshmen, another for sophomores with parameters:

	Freshmen	Sophomores
Mean	$\mu_a = 61.2$	$\mu_b = 73.7$
S.D.	$\sigma_a = 8.52$	$\sigma_b = 9.36$

(*a*) What is the z score, z_a, with respect to the freshman distribution of a score, X_i, which has a z score of $z_b = +.89$ with respect to the sophomore distribution?

Note: Since

$$X_i = \mu_a + z_a\sigma_a$$
$$= \mu_b + z_b\sigma_b$$

therefore, $\mu_a + z_a\sigma_a = \mu_b + z_b\sigma_b$
and, solving for z_a,

$$z_a = \frac{\mu_b - \mu_a}{\sigma_a} + \left(\frac{\sigma_b}{\sigma_a}\right) z_b$$

13. Discuss the implications of Ex. 12 for the establishment of test norms.

14. Prove that the value of $X_b(a)$, a score of the distribution of a which will have the same z score in distribution a as X_b has in distribution b, is

$$X_b(a) = (\sigma_a/\sigma_b)X_b - [(\sigma_a/\sigma_b)\mu_b - \mu_a]$$

15. By means of Ex. 14 find, from the information of Ex. 12, a sophomore score in the *freshman distribution* equivalent to an actual sophomore raw score of 80. In what respects would your result have the same interpretations for sophomores as an identical "freshmen raw score" would have for freshmen?

REFERENCES

1. Dixon, Wilfrid J., and Frank J. Massey, *Introduction to Statistical Analysis*, New York: McGraw-Hill Book Co., 1951, Chapter 5.
2. Freund, John E., *Modern Elementary Statistics*, New York, Prentice-Hall, 1952, Chapter 6.
3. Guilford, Joy P., *Fundamental Statistics in Psychology and Education*, Second Ed., New York, McGraw-Hill Book Co., 1950. Chapters 12 and 19.
4. Gulliksen, Harold, *Theory of Mental Tests*, New York, John Wiley and Sons, 1950, Chapters 18, 19, and 21.
5. Johnson, Palmer O., and Robert W. B. Jackson. *Introduction to Statistical Methods*, New York, Prentice-Hall, 1953, Chapters 8 and 9.
6. Lindquist, Everet F., *A First Course in Statistics*, Revised Ed., Boston, Houghton Mifflin Co., 1942, Chapter 7.
7. Lindquist, Everet F., editor, *Educational Measurement*, Washington, D. C., American Council on Education, 1951. John C. Flanagan, "Units, Scores, and Norms," pp. 695-763.
8. Walker, Helen M., *Elementary Statistical Methods*, New York, Henry Holt and Co., 1943, Chapter 11.
9. Wilks, Samuel S., *Elementary Statistical Analysis*, Princeton, N. J., Princeton University Press, 1949, Chapter 8.

CHAPTER 7

The Sampling Distribution of the Mean

We have seen that an investigator may have interest in a *parameter*, that is, some statistical measure descriptive of the distribution of a *universe*. We have seen furthermore that most often the investigator does not have the data from an entire universe. In Chapters 3 and 4 measures of central value and dispersion were discussed, *as if* such measures were derived from an entire universe. From a sample we can compute descriptive measures—that is, *statistics*—from which we may wish to make inferences regarding the universe. In Chapters 5 and 6 we saw that there are *theoretical* distributions, derived mathematically on assumptions regarding the variables with which they deal. If a theoretical distribution be assumed to represent the universe, it can be used to supplement information obtainable from sample data, the value of such supplementary information depending on how near the theoretical is to the true distribution.

As an example, a measure in which there is frequent interest is the mean. There are two means which the investigator usually must consider, whether he succeeds in actually computing both of them or not:

$$\text{The actual universe mean, } \mu = \frac{1}{N}\sum_{i=1}^{N} X_i$$

$$\text{The sample mean, } \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Some suggestion of the use of the mathematical model or theoretical distribution in relating *statistic* to *parameter* and *sample* to *universe* was covered in Section 5.6 in our discussion of the binomial distribution. It is the purpose of this chapter to take us one step further along the road of relating *statistic* to *parameter*, with particular reference to the normal distribution.

7.1 SOME BASIC THEOREMS CONCERNING THE MEAN

The logical justification for the most important uses to which we put the normal distribution depends upon proofs beyond the scope of this book. Fortunately, the theorems themselves are not complex, and their correctness may easily be tested empirically.

The theory of the *sampling distribution*, which we are about to discuss, assumes *randomness* of sampling as defined in Section 5.7. In this discussion we will make use of a device. The sampling distribution of a statistic (such as $\bar{X}$) is the distribution of the values of the statistic computed from an infinitely large number of (or all possible) samples of the same size, n, drawn from the universe.

Thus a sample may be viewed as a sequence of n independent repetitions of an operation or a measurement from a universe of such operations or measurements. In turn, any *function* of such sample values X_1, $X_2, \cdots, X_n$, may be shown to be a random variable. In this way, the statistic, which varies from sample to sample, is different from the parameter, which is a *constant*.

By mathematical reasoning, characteristics of the sampling distribution of a mean, a median, a variance, a percentile, or any other statistic, can be expressed in terms of the parameters of the distribution of the universe. For the present, we restrict our remarks to infinitely large populations, though this is not essential to all aspects of our discussion. This is comparable to sampling chips, on which there are numbers, from a bowl containing such chips, but replacing each sample chip before the next one is drawn. Most of the statistical techniques and the statistical theory with which the educator deals is concerned only with sampling from very large universes, or sampling as though *with replacement*. We shall therefore not concern ourselves for the time being with sampling from "finite" populations, and particularly not with sampling from small universes without replacement.

The following principles regarding the *sampling distribution* of a mean may now be stated:

THEOREM (a). *The mean of means of all possible random samples of size n drawn from a population equals the mean of the population.* That is, the mean of the sampling distribution of sample means is the population mean. Using "expected values," which represent means of theoretical distributions,

$$E(\bar{X}) = \mu$$

where $\bar{X}$ is a sample mean, and μ is the mean of the universe from which the sample is drawn.

THEOREM (*b*). *The variance of the sampling distribution of means of samples of size n is equal to* $1/n$ *times the variance of the population.* In symbols,

$$\sigma^2_{\bar{X}} = \sigma^2/n \tag{7.1}$$

where $\sigma^2_{\bar{X}}$ equals the variance of sample means;
σ^2 equals the variance of the population;
and n equals the size of sample.

THEOREM (*c*). *The sampling distribution of means of samples of size n from a normal population is itself normal.*

THEOREM (*d*). *The sampling distribution of means of size n has a mean* μ *and variance* σ^2/n, *where* μ *and* σ^2 *are respectively, the mean and variance of the population sampled, and, for a wide variety of nonnormal populations, approaches a normal distribution as n becomes increasingly large.*

The last two theorems place strategic emphasis upon the normal distribution in sampling theory. Examples and demonstrations can be used to show the practical significance of these theorems.

Comprehension of these theorems is one of the important rewards in the study of statistics. Equation 7.1 is an exceedingly useful tool, and an understanding of it is one of the chief distinctions between statistically literate and statistically illiterate persons. The layman usually judges the dependability of a sample mean in terms of the *proportion* of the population which is sampled. But, according to our theorem, the dependability of a sample mean is primarily a function of other factors. In large populations, according to this theorem, the variance of sample means is directly related to the variance of the population, and inversely related to the number of cases in the sample.

Equation 7.1 may be expressed as a formula for the *standard error*, $\sigma_{\bar{X}}$, of a mean $\bar{X}$, in terms of the universe standard deviation. By taking the square root of equation 7.1 we may write

$$\sigma_{\bar{X}} = \sigma/\sqrt{n} \tag{7.2}$$

The term *standard error* is commonly used for $\sigma_{\bar{X}}$ as a measure of the variability of sample means. Of course, error does not mean mistake, rather *sampling* error or, better still, sampling variation, in the sense in which we have used error in previous discussions dealing with sampling.

It is emphasized that equations 7.1 and 7.2 are applicable only when the *universe variance*, or standard deviation, *is known*.

We shall now seek some assurance of the correctness of Theorem (*b*). The implications are these:

(1) Given a universe with variance σ^2, the *larger* the n, the size of the sample, the *smaller* the variance of the sample mean.

(2) Given a sample of size n, the larger the variance of the population sampled, the larger the variance of the sample mean.

7.2 A DEMONSTRATION OF THE VARIANCE OF A SAMPLE MEAN

A common device in classes in elementary statistics is the experiment based upon drawing a large number of samples of a given size from a normal population to test Theorems *a* and *b* above. Results of such experiments appear in a number of textbooks.[1] A suggested experiment of this type appears as one of the exercises at the end of this chapter.

Since the chief power of the ideas we have discussed in this chapter to this point draws upon Theorem *d*, we shall examine some nonnormal populations. If equation 7.1 regarding the variance of a mean may be used with good approximation even for samples from nonnormal universes, that is of great importance. It would mean that, by use of a formula, we can compute the variance of the distribution of sample means, without the laborious task of drawing many samples in order to find the distribution of the sample means.

Our first experiment will deal with a very simple universe. It is a universe of the four values 1, 2, 3, and 4. Imagine, for instance, a bowl in which there are placed four chips, one of which is marked with the number 1, the second with 2, the third with 3, and the fourth with 4. Imagine, furthermore, that we are interested in the distribution of means of samples of size 2; that is, we shall draw many samples of two individuals from this universe, drawing one chip at a time with replacement.[2] Drawing with replacement, that is drawing one, replacing it and drawing another, is equivalent to drawing from a very, very large bowl in which there are many chips marked 1 and an equal number of chips marked 2, 3, and 4. The distribution of this universe is

X	$P(X)$
4	.25
3	.25
2	.25
1	.25

This is a probability distribution. It is made up just as we made up theoretical distributions of the binomial in Chapter 5. We could plot a histogram of this distribution as we did with binomials, and all the columns

[1] See, for example reference 2, pp. 40-42, and reference 5, pp. 33-37.

[2] This experiment is adopted from W. Edwards Deming, reference 1.

over the possible X values would be the same height, .25. In other words, this is a discrete distribution which does not have the bell-shaped appearance of the normal curve or of the binomial. It is therefore decidedly *nonnormal*.

The mean of this universe can be computed according to equation 5.5:

$$\begin{aligned}\mu = E(X) &= \Sigma XP(X)\\ &= (.25)(1 + 2 + 3 + 4)\\ &= 2.5\end{aligned}$$

The variance of this theoretical distribution can be computed as follows:

$$\begin{aligned}E(X^2) &= \Sigma X^2P(X)\\ &= (.25)(1 + 4 + 9 + 16)\\ &= 7.5\end{aligned}$$

Since, from equation 5.7

$$\sigma^2 = E(X^2) - (EX)^2$$

we find the variance

$$\sigma^2 = 7.50 - 6.25 = 1.25$$

We now have from a theoretical distribution the theoretical mean, μ, and the theoretical variance, σ^2, of the total universe.

According to Theorem *a* above, the mean of all possible samples of size 2 from this universe would be the mean of the universe, 2.5. Moreover, according to Theorem *d*, the distribution of all possible sample means should tend to the normal, and the variance of this distribution of means should be σ^2/n. We thus have the following theoretical information about the distribution of the sample means:

$$E(\bar{X}) = \mu = 2.5$$

and

$$\sigma_{\bar{X}}^2 = \sigma^2/n = 1.25/2 = .625$$

where $E(\bar{X})$ is the mean of the universe, 2.5, which we expect for $\bar{\bar{X}}$, the mean of all possible sample means; and where $\sigma_{\bar{X}}^2$ is the variance of all possible means of samples of 2. The only other bit of theoretical information which the theorems tell us concerns the character of the distribution of the sample means. As pointed out above, the distribution of raw scores is itself flat. According to Theorem *d*, the distribution of means should be nearer to the normal, that is, there should be higher frequencies near the mean value, 2.5, and lower frequencies at each end of the distribution.

We are now ready to seek empirical support for these theorems applied to our particular distribution. Above we have the theoretical information, and in Table 7.1 are all possible samples of size 2 which can be drawn *with replacement from this universe*. The first possible sample listed is the result of drawing the 1 chip the first time and also the second time. The next one is the sample consisting of the 1 chip and the 2 chip, and so on. The 16 possible samples are shown in the first column of Table 7.1.

TABLE 7.1

ALL POSSIBLE SAMPLES OF SIZE 2 FROM POPULATION 1, 2, 3, 4, FOR WHICH $\mu = 2.5$ AND $\sigma^2 = 1.25$ (WITH REPLACEMENT)

Sample Pair	Sample Mean $\bar{X}$	$(\bar{X} - \bar{\bar{X}})^2$	$\Sigma(X - \bar{X})^2$ or s^2
(1)	(2)	(3)	(4)
1–1*	1.0	2.25	.00
1–2	1.5	1.00	.50
1–3	2.0	.25	2.00
1–4	2.5	.00	4.50
2–1	1.5	1.00	.50
2–2*	2.0	.25	.00
2–3	2.5	.00	.50
2–4	3.0	.25	2.00
3–1	2.0	.25	2.00
3–2	2.5	.00	.50
3–3*	3.0	.25	.00
3–4	3.5	1.00	.50
4–1	2.5	.00	4.50
4–2	3.0	.25	2.00
4–3	3.5	1.00	.50
4–4*.	4.0	2.25	.00
Totals	40.0	10.00	20.00
Averages	2.5	.625	1.25

* Samples which would not occur in sampling *without* replacement.

The mean of the two members of each sample, shown in column 1, is reported in column 2. The sum of these 16 means is 40. Consequently, the mean of the means is $\bar{\bar{X}} = 2.5$, confirming Theorem *a*.

Let us now turn to the variance of these means. The *squared* deviation of each mean from the general mean, 2.5, is shown in column 3. The sum of these 16 squared deviations is 10, and the mean square deviation or variance of the means is therefore .625. In symbols:

$$\sigma^2_{\bar{X}} = \Sigma(\bar{X} - \bar{\bar{X}})^2/k = 10/16 = .625$$

which checks perfectly with the theoretical value.

The frequency distribution of the 16 possible means is

$\bar{X}$	f
4.0	1
3.5	2
3.0	3
2.5	4
2.0	3
1.5	2
1.0	1

It is clear that, although the parent population had the same frequency for each value of X, the distribution of the means is nearer to the normal.

We shall refer later to the fourth column of Table 7.1, which shows how *variances* behave in samples.

7.3 SAMPLING FROM A VERY SKEWED DISTRIBUTION

The above demonstration has the advantage of being simple, the numerical computations needed to verify the theory being easy. It involves one type of "lack of normality" of the parent population. The universe is decidedly platykurtic.

There is some advantage in seeing what actual data from a real distribution, not a theoretical one, will show us regarding our theorems. One kind of absence of normality which may lead to poor results from the formula for the variance of sample means is *skewness*. A skewed distribution is shown in Table 7.2. It is the distribution of enrollments in high schools in the State of Illinois as reported in the 1951-1952 Illinois School Directory. It is hardly necessary to compute indices of skewness for this distribution. A glance at Table 7.2 shows that 450, considerably over half, of the high schools are in the bottom two-class intervals, with enrollments less than 200. The stringing out of the right-hand tail of the distribution is partly concealed by the use of larger intervals for enrollments over 1,000, to avoid a lengthy tabulation. The last four intervals are all greater than the interval size of 100 used for the others.

According to Theorem *d*, we should not expect the distribution of means to follow the normal distribution unless n is very large, though the formula for the variance of the means should correspond with the actual computed variance within errors of random sampling. Therefore, it is informative to see how nearly a sample of means from this universe will approach the normal distribution.

TABLE 7.2

FREQUENCY TABLE OF ILLINOIS HIGH-SCHOOL ENROLLMENTS

Enrollment	*f*
3000 and over	3
2000–2999	8
1500–1999	5
1000–1499	26
900– 999	5
800– 899	8
700– 799	12
600– 699	14
500– 599	16
400– 499	32
300– 399	50
200– 299	82
100– 199	243
0– 99	207
Total	711

In Table 7.3 are the distribution of 100 sample means, each based on 10 items drawn *with replacement* from this distribution. The samples were drawn in this way:

Each of the 711 high schools was assigned a serial number. The first one was 001, the next one was 002. Next a table of random numbers similar to the one shown in Appendix B was entered at an arbitrary starting point. Three columns of the random numbers were used. Beginning at the random place in the table, the three digits in the starting line in the three columns were recorded, then the next line, and so on until ten random numbers of three digits had been entered. These random numbers were then used as serial numbers to identify the ten schools to appear in the first *random* sample of ten. The table of random numbers was then used in a similar way for the next ten by continuing down the three columns. When the bottom of the table was reached, the next three columns were used and so on until 100 sets, each of 10 random numbers of three digits each had been entered. As with the first sample,

TABLE 7.3

DISTRIBUTION OF MEANS FROM RANDOM SAMPLES OF HIGH-SCHOOL ENROLLMENTS, $n = 10$ AND $n = 20$

Enrollment	Frequency	
	100 samples $n = 10$	50 samples $n = 20$
750–799	1	
700–749	1	
650–699	2	1
600–649	4	
550–599	1	
500–549	3	
450–499	1	5
400–449	7	1
350–399	5	7
300–349	12	9
250–299	16	3
200–249	15	14
150–199	21	7
100–149	9	3
50– 99	2	
Total	100	50

the random numbers identified the high schools (by serial numbers which had been previously assigned), constituting each of the 100 samples.

The second column in Table 7.3 shows a distribution of 50 samples, with $n = 20$. This was made by combining pairs of samples of 10 schools. For instance, the first and the second were combined to make up a sample of 20, the third and fourth another sample of 20, and so on.

The data in Table 7.3 give us some notion of the kind of distribution resulting from sampling the very skewed distribution of Table 7.2. The 100 samples of size 10 distribute themselves in a positively skewed manner, but clearly much less skewed than the parent distribution. Note, for instance, that the highest frequencies of sample means are near the mean of the parent distribution ($\mu = 289.5$), and lower frequencies occur on the negative end as well as the positive end of the distribution. Nevertheless, there is still a much greater span of variation of means above the central interval of 250 to 299 than below it.

A second observation is that the spread of the means is less when the sample size is 20, in accord with Theorem *b*. Also the distribution appears skewed. There is, of course, sampling variation in the sample of means, and we do not have many samples, either the 100 of size 10 or the 50 of size 20. It is, therefore, hardly worth the trouble to compute measures of skewness, kurtosis, and so on. By observation, however, we see that with $n = 20$, though we come nearer the normal curve than the parent distribution, we still have skewness in the distribution of means. Let us compare the theoretical standard deviation and means with the observed standard deviation and means to get a feel of the extent to which we can make use of Theorem *d* above.

The theoretical values are computed in the following. By the gross score method (that is, not using grouped data), the mean of the universe, the data of Table 7.2, was found to be 289.5 and the standard deviation was found to be 413.9. By means of equation 7.2 the value of the standard deviation of sample means of size 10 should be

$$\sigma_{\bar{X}} = \sigma/\sqrt{10} = 413.9/3.162 = 130.9$$

Similarly, the theoretical standard deviation of means of samples of size 20 is 92.6.

Our experimental values are derived by actually computing the mean of the sample means and the standard deviation of the sample means shown in the distributions of Table 7.3. The result is summarized in Table 7.4.

TABLE 7.4

COMPARISON OF MEANS OF SAMPLE MEANS, OBSERVED STANDARD ERROR OF MEANS AND UNIVERSE VALUES, TWO SAMPLE SIZES FROM ENROLLMENTS OF ILLINOIS HIGH SCHOOLS

Item	Theoretical, by Formula	Observed by Computation
100 samples, $n = 10$		
Mean of means	$\mu = 289.5$	$\bar{\bar{X}} = 293.3$
S.D. of means	$\sigma_{\bar{X}} = 130.9$	$s_{\bar{X}} = 149.8$
50 samples, $n = 20$		
Mean of means	$\mu = 289.5$	$\bar{\bar{X}} = 293.3$
S.D. of means	$\sigma_{\bar{X}} = 92.6$	$s_{\bar{X}} = 112.1$

We find that the mean of the sample means is fairly close to the universe value, that is, the expected value. In fact, the difference between the observed value 293.3 and the universe value 289.5 resulted from the use of a finite number of samples. In either case the information is supplied, at least in round numbers, that the average size of high school is about

290. This we may consider a reasonable check on our anticipation that the mean of means would come close to the universe mean.

A different story is told by the standard deviation of the means. For the 100 samples of size 10, the theoretical value is 130.9, compared with the observed value of 149.8. In part this is due to the fact that our experiment included a finite number of samples, but in part it results from the skewness of the distribution. How we are to judge this result depends upon the use to which it may be put. In other words, if a rough idea of the standard deviation of sample means was desired, the figure 131 by formula might be considered acceptable if the true standard deviation was 150.

There is likewise a difference between the theoretical and the observed variances for samples of size 20. Part of this difference is due to the fact that we have only 50 sample means—not all possible sample means of size 20. Even so, if it is sufficient to know that the standard error is something around 100, our experimental results show that the theoretical value is adequate.

According to the theory, if we were to take a large number of samples of size greater than 20, say with $n = 50$ or $n = 100$, the variances of the means would be less. The larger the number of cases in the sample, the smaller the variance. This would hold whether dealing with the theoretical formula or dealing with the experimental values as in Table 7.4.

7.4 TESTING STATISTICAL HYPOTHESES

We have now learned something about the sampling distribution of one *statistic*, the mean. We shall learn later about sampling distributions of other statistics. Since many types of problems are concerned with the mean and since the sampling distribution of the mean is among the simplest, we shall discuss sampling applications of the mean in some detail.

One feature of the sampling distribution of the mean, as we have seen, is that it follows the normal distribution (1) exactly if the universe is normal, and (2) otherwise approximately if n is large. In symbolic terms the ratio $z = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ is normally distributed with mean zero and unit variance when the parent population is normal. This ratio will be recognized as expressing the deviation of the sample mean, $\bar{X}$, from the population mean, μ (or *true* mean as it is sometimes called), in *standard deviation* units.

There are many other statistics for which the normal distribution is a suitable approximate sampling distribution. This is one of the chief reasons why we spent some time on its study in Chapter 6. For the present we will exploit our newly acquired knowledge of the distribution of sample means to see how the normal distribution is used in *statistical inference.*

As stated in Chapter 5, the systems of statistical method employed in making statements about universes from samples is called *statistical inference.* One of the devices for doing this is *testing a statistical hypothesis.* A statistical hypothesis is an assumption concerning a *parameter.* If θ represents some parameter of a universe, $H : \theta = \theta_0$ is one method of stating the hypothesis that θ has some specified value θ_0. The hypothesis that the mean of the universe is 50 would be written $H : \mu = 50$.

A test of a statistical hypothesis is a procedure for deciding whether to *reject* a hypothesis. Only from a complete enumeration or measurement of a population could we *prove* a hypothesis. A test of a statistical hypothesis is made only in terms of a probability statement. In order to have the necessary "probability" information we must know or be able to estimate satisfactorily the probability distribution of the *statistic* (that is, a value computed from a sample), this distribution being the basis on which the hypothesis can be tested.

A *critical region* or *rejection region*—part of all possible values of a sample statistic—is established for which we will *reject* the hypothesis if the sample statistic is contained within it. This critical region is chosen in such a way that such values would be expected to occur by chance *rarely* if the hypothesis tested is true. How rarely, depends upon the *risk* we are willing to take of rejecting a hypothesis when it is true. Most commonly statisticians define "rarely" as $P = .01$ or $P = .05$. The probability, α (usually .01 or .05), is used to determine the critical region. It is called the *level of significance.* If, on a given hypothesis, the probability of occurrence of the sample value is equal to or less than α, the result is said to be *significant.* This is a brief way of saying that a result has been observed which would be rare if the hypothesis being tested were in fact true.

7.5 HYPOTHESES REGARDING THE POPULATION MEAN—σ KNOWN

The procedure will be illustrated with a problem. We will discuss the theory of the solution of this problem in some detail, as the reasoning behind the testing of statistical hypotheses is very important, particularly in more complex problems. Moreover, the procedure is applicable to statistical tests concerning parameters other than the mean. In practice a statistician will test a hypothesis in very simple terms by "finding out whether or not it is *significant*," as suggested above, or simply by inquiring as to the probability of the occurrence of a sample value on the basis of some assumption regarding the universe from which it is drawn. Our objective, however, is to understand the reasoning basic to the conventions which have been established in the field of statistics. We shall, therefore, examine the explanation. In the beginning it will seem to be roundabout, but as we follow the steps through we should be able to see why it is that the roundabout reasoning is considered important.

A test administered to entering freshmen has an established mean μ of 73 and a σ of 12. A sample of 16 students is found to have a mean score, $\bar{X} = 66$. We wish to determine whether or not this may be considered a random sample from the population for which $\mu = 73$ and $\sigma = 12$.

In this case the population variance is known (or we are willing to assume it). Ou1 hypothesis is $H : \mu = 73$.

The procedure in testing the hypothesis is as follows:

Step 1. *Choose a Level of Significance.* We decide to use .05.

Step 2. *Determine Distribution to Use.* This concerns the sampling distribution of the mean. We assume that the population is normal, as is approximately the case with the test in question. Hence the sampling distribution is also nearly normal and the statistic $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ is nearly normally distributed with unit variance and, if the hypothesis is true, with mean zero.

Step 3. *Establish Critical Region.* We select the critical region as the two extreme parts of the distribution, each consisting of the 2.5 percent of largest deviations from μ. The critical region, or rejection region, is the pair of intervals of $z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ for which $z < z_{.025}$ and $z > z_{.975}$, where $z_{.025}$ and $z_{.975}$ are standard scores such that $\int_{-\infty}^{z_{.025}} = \int_{z_{.975}}^{\infty} = .025$. Looking in Appendix D, we find that $z_{.975} = 1.96$, and, of course, $z_{.025} = -1.96$. Hence,

$$P\left(-1.96 < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 1.96\right) = .95$$

and

$$P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < -1.96\right) = P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} > 1.96\right) = .025.$$

Figure 7.1. shows the critical region thus chosen, consisting of the two tails of the distribution of sample means. The variance of this distribution is the variance of means of samples of the given size. The central value is μ_0, the assigned value involved in our hypothesis. In our particular example

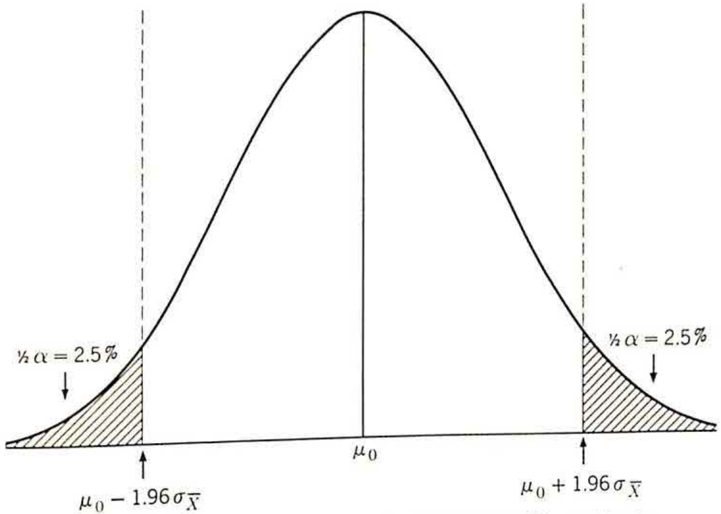


FIG. 7.1. Critical region for two-tailed test of hypothesis H: $\mu = \mu_0$ at 5 percent level.

the standard deviation of the sampling distribution would be $\sigma_{\bar{X}} = (\sigma/\sqrt{n}) = (12/4) = 3$. In score units the critical region is the set of values more than $1.96\sigma_{\bar{X}}$ units *below* μ, $[73 - (3)(1.96)]$, and the set of values more than $1.96\sigma_{\bar{X}}$ units *above* μ, $[73 + (3)(1.96)]$. That is, we will reject the hypothesis if $\bar{X} < 67.12$ or $\bar{X} > 78.88$.

Step 4. *Compare Sample Statistic and Critical Region.* If we convert the boundaries of the critical region into score units as in the preceding sentence, we accept or reject the hypothesis directly on the basis of $\bar{X}$. In our particular problem $\bar{X} = 66$. It is less than 67.12 and is therefore in the rejection region. The test may be made directly in terms of standard scores. We might have computed $z = (66 - 73)/3 = -2.33$. Since this is less than -1.96, it falls in the rejection region, giving us the same result. The probability of a deviation from the mean of as much as (or more than) $2.33z$, in either direction, positive or negative, may be found in the table of Appendix D. It is $2\int_{2.33}^{\infty} = 2(.01) = .02$.

Now actually what we have done up to this point is to find out that if the universe from which we have sampled has a mean of 73, it is unlikely that sampling alone would produce the mean of 66 which we observed.

Our impulse is to conclude that the hypothesis is a very improbable one. However, in modern statistical theory we say that we do not know anything about the probability of the value of the mean of our universe. Whatever it may be, it is some specific value. The probability that it is any specified value whatsoever is either zero or one. We can say, however, that our hypothesis was one for which *the observed sample mean would be so unlikely to occur*, that we are led to conclude that the hypothesis is false.

7.6 ONE-TAILED AND TWO-TAILED TESTS— σ KNOWN

The foregoing is an illustration of a "two-tailed" test of a mean. Our hypothesis was such that we would reject a very large positive *or a very large negative* observed deviation from the hypothetical population mean, μ_0. We were interested in knowing whether the sample mean could have been a random variation from the hypothetical population mean *in either direction.* In making the test, we were asking whether $| z |$, that is, the magnitude of z without regard to sign, was greater than 1.96.

This is the procedure to follow unless there is some reason for considering only positive or only negative differences between $\bar{X}$ and μ_0. Consider a methods experiment involving 25 pupils who are to be instructed under a new system of teaching an English course. The mean of previous students on a.test has been 70, and σ has been 20. We plan to give this same test to the experimental group, a sample of 25 students. We will not consider the possibility that we have a *poorer* method, because our interest is in finding better methods.

As previously, we do not, however, directly test the hypothesis, that the method is *better* than the usual instruction. The hypothesis to test in this case is that the population, *whatever it is*, from which the sample of 25 cases was drawn, *is one whose mean is not greater than* 70. In other words, we are to test $H : \mu \leq \mu_0$, the hypothesis that the unknown mean, μ, of this theoretical population, is *equal to or less than* μ_0, some specified value. The specified value, μ_0, in this case is 70.

Suppose that at the end of the experiment the measure we use produces a mean, $\bar{X} = 65$, for the sample of 25 subjects. This outcome is in agreement with the hypothesis, giving no indication that the new method has a higher mean *than the conventional method.* In fact, all values of $\bar{X} < \mu_0$ would agree with the hypothesis. In this case we do not include in the rejection region any of the possible values of $\bar{X}$ which are below, that is, less than μ_0.

Now suppose instead that we find $\bar{X}$ to be 77. Then we find that $z = 1.75$. The rejection region, at the 5 percent level, assuming the normal distribution, is shown graphically in Fig. 7.2. It is $z > z_{.95} = 1.65$. The observed z is, therefore, in the rejection region. There is less than a 5 percent chance that the sample mean of 77 could occur if its population mean were 70 *or less*. Therefore, we reject the hypothesis.

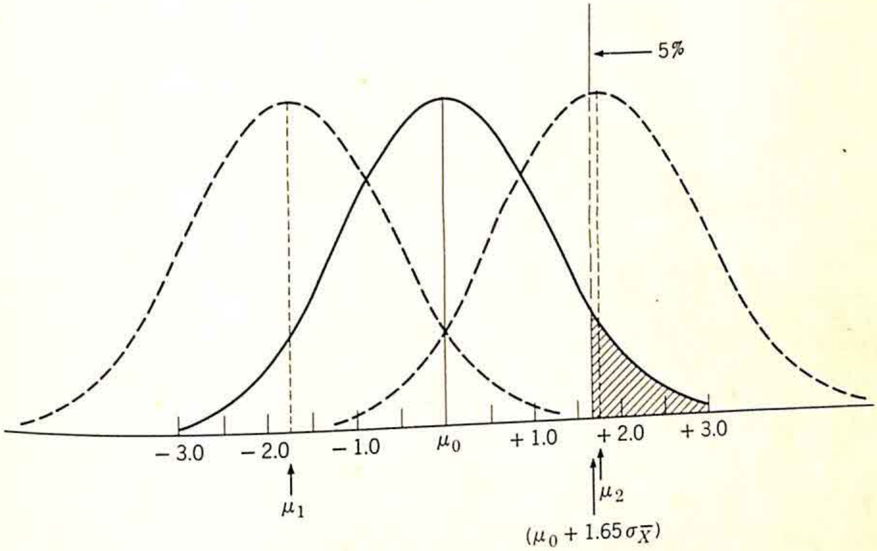


FIG. 7.2. Diagram for test of one-tailed hypothesis H: $\mu \leq \mu_0$ and alternative hypotheses.

Figure 7.2 should assist in understanding some of the consequences of deciding in this way to reject the hypothesis, and should illustrate some of the differences between the one-tailed and the two-tailed tests. In the example that we have just discussed, μ_0 was 70. The observed $\bar{X}$ was 77, and it was seen to be $\mu_0 + 1.75\sigma_{\bar{X}}$. Thus it fell in the rejection region. Therefore we decided that there was slight chance of observing this value in sampling a distribution for which μ_0 is the mean, and we concluded that our sample came from a distribution whose mean is larger than μ_0. The solid-lined curve in Fig. 7.2 is the probability distribution used for the test. The dotted lines represent the distributions for two *alternative* hypotheses, $H : \mu = \mu_1$, and $H : \mu = \mu_2$.

It is to be noted that we have taken a calculated *risk*, α, when we reject the hypothesis, that it could be one of the rare observed statistics falling in the rejection region *when in fact the hypothesis is true*. This is sometimes called the Type I error in testing a statistical hypothesis, the probability of which, α, is the probability that the hypothesis will be rejected *when it is true*. This we control, before making the statistical test, by establishing whatever value we wish for α. It was .05 in our example.

Had we chosen to be more (or less) willing to make this kind of error, we would have used another value of α. It must be remembered that *rejecting* the hypothesis would favor adoption of the new method. Note that the hypothesis tested was that the new method was no better than the current practice. If adopting the innovation would be costly or if for any reason we were anxious to reduce the chance of Type I error, we might choose .01 or even .001 for α. If we had reasons other than the outcome of the test for adopting the new practice, we might choose .10 or .20 or some other higher value for α.

There are other probabilities to take into account. They are suggested in Fig. 7.2. One of them is the probability of *accepting* (at least not rejecting) the hypothesis *when it is in fact true*. In Fig. 7.2 this is represented by the area under the solid curve to the left of the critical region, which is $1 - .05 = .95$. In general the probability of accepting a hypothesis when true is $1 - \alpha$.

Now let us look at the conceivable position μ_1 of the true mean score for the new method. It illustrates a value of μ less than the specified value μ_0 (70 in our example). In the one-tailed test there is very little chance of rejecting the hypothesis in sampling a population whose mean has this value, since only a small tip of the tail of the distribution of sample means reaches the rejection region for the hypothesis which we have established. In other words, if the experimental method is on the whole really less effective, a single sample mean tested in this way is not likely to be high enough to prove "significant" and hence to cause us to favor the (wrong) conclusion that the new method has a *higher* population mean.

7.7 THE POWER OF A TEST

In more complex types of statistical tests some such reasoning as the foregoing is involved in choosing the "best statistical tests." One requirement of a good statistical test is *a high probability of rejecting a hypothesis if false*. This is called the *power of a test*, $(1 - \beta)$. Unlike α, it is not a single value which the investigator may establish at whatever level he pleases for a given sample size n. The value of $(1 - \beta)$ depends upon the value of μ; $(\mu \neq \mu_0)$; μ being unknown, we can only speculate as to the values which it may take.

One conceivable value of μ is μ_2, referrring again to Fig. 7.2. The part of the distribution of the sample means about μ_2 which is in the rejection region is obviously considerably more than 5 percent. There is thus a substantial probability that this value of μ would give us a sample mean, $\bar{X}$, that would lead us to reject the hypothesis that $\mu \leq \mu_0$. This probability is the power of the test, $1 - \beta$, for the particular value $\mu = \mu_2$.

The area β in the μ_2 distribution, to the left of $\mu_0 + 1.65\sigma_{\bar{X}}$ and so not in the critical region, is the probability of accepting the *false hypothesis* that $\mu \leq \mu_0$. This is sometimes called the Type II error.

In a similar manner we can determine probabilities β and $1 - \beta$ for either one-sided or two-sided tests at any level α for any number of conceivable values of μ other than μ_0, the value specified in a hypothesis. The results would permit a comparison of the power of the two types of tests (for various conceivable values of μ). It would be found that this may be said for the *one-sided test*:

(1) It is a more powerful test (will reject more false hypotheses) than the two-sided test for all values of $\mu > \mu_0$ (or $\mu < \mu_0$ in a left-handed test).

(2) It has little or no power to reject the hypothesis $H : \mu = \mu_0$, when in fact μ is less than μ_0. Its power is less than α.

Regarding the two-sided test, we may state the following:

(1) It has more power, that is, a greater probability of rejecting a *false* hypothesis $H : \mu = \mu_0$ whether μ is greater or less than μ_0.

(2) It is less powerful than a one-tailed test when considering alternative values of μ on the same side of the distribution as the one-tailed rejection region. It is of course obvious that there are two one-sided tests for the normal distribution, depending on whether we are to reject statistics *greater than* the specified value of the parameter or *less than* the specified value of the parameter.

7.8 CONFIDENCE INTERVALS FOR POPULATION MEAN

The objective of a statistical investigation is not always *experimental* in nature. The purpose may be merely *descriptive*, that is, the investigator would like to know from a sample as much as he can about a parameter (or parameters) of a population. Concerning $\bar{X}$, a sample mean, we know that it is an unbiased estimate of μ, unbiased because the average of all possible sample means will be μ. But if we have only a single sample mean, we may wish to know its relationship to μ.

To illustrate the problem we shall use the approximately normally distributed California Test scores in Appendix A. We are to compute a mean from a sample of 10 scores from this distribution. Let us now imagine that we are to use this as an estimate of μ and that μ is unknown. If σ is known, we may compute $\sigma_{\bar{X}}$. In this example, $\sigma = 13.19$. Hence $\sigma_{\bar{X}} = 13.19/\sqrt{10} = 4.17$.

We may determine the probability that the variable $(\mu - \bar{X})/\sigma_{\bar{X}}$ will be contained within certain limits. For instance, the probability is .95 that $(\mu - \bar{X})/\sigma_{\bar{X}}$ will be between $z_{.025} = -1.96$ and $z_{.975} = 1.96$. This may be expressed as follows:

$$P[-1.96\sigma_{\bar{X}} < (\mu - \bar{X}) < 1.96\sigma_{\bar{X}}] = .95$$

Adding $\bar{X}$ to each term of the inequality, we have

$$P(\bar{X} - 1.96\sigma_{\bar{X}} < \mu < \bar{X} + 1.96\sigma_{\bar{X}}) = .95$$

It is to be noted that $\bar{X}$ is the variable from sample to sample, not μ. Hence, the meaning of the foregoing equation is that for 95 percent of samples, $\bar{X} \pm 1.96\sigma_{\bar{X}}$ will contain μ. We could have selected $z_{.995} = 2.58$ and the probability would have been .99, or we could have selected the z corresponding to any other probability.

In our example, the limits $\bar{X} - (1.96)(4.17)$ and $\bar{X} + (1.96)(4.17)$ are the *confidence limits* for a 95 percent *confidence interval.* One set of 95 percent confidence limits for an actual sample of the California Test scores, for which $\bar{X}$ is 69.0, is $69.0 - 8.17 = 60.8$ and $69.0 + 8.17 = 77.2$. Another sample of 10 with $\bar{X} = 73.2$ has limits 65.0 and 81.3; a third sample with $\bar{X} = 71.9$ has limits of 63.7 and 80.1.

We do not *know* that any single one of the three intervals will contain μ, but we have considerable "confidence" in that possibility, since about 95 of 100 such sample intervals *will* contain μ. By the same token about 5 of 100 intervals *will not* contain μ. The level of assurance that μ will be contained in a single sample interval is called the *confidence coefficient.* In the above case it is .95. It does not *guarantee* that we have a good estimate of μ or even that μ is within the interval. It does tell us that if we follow this method, 95 per cent of such confidence intervals will include the value of the parameter.

The vertical lines in Fig. 7.3 represent several confidence intervals computed from sample means. Most of them intersect the horizontal line, representing μ. Our discussion has assumed random sampling. Under conditions of random sampling, as we have noted earlier, the mean of sample means is equivalent to the population mean. In Fig. 7.3 the mid-points of the intervals represent the sample means. They are distributed about μ as the central value. If sampling is *not* random, the horizontal line of the mean of sample means might not coincide with the line representing μ. The distance between the two lines would represent "bias." In a diagram of intervals from *biased* samples, all intervals would be shifted up or down by the amount of this bias. The result would be a decrease in the number of intervals intersecting μ. The confidence coefficient would be a false index of the assurance with which an interval would contain μ. The confidence interval is based on the sampling distribution of the mean as a *random variable*. This holds only under conditions of random sampling.

Exercise 2 at the end of this chapter suggests a class experiment which may be used to demonstrate the idea of confidence intervals.

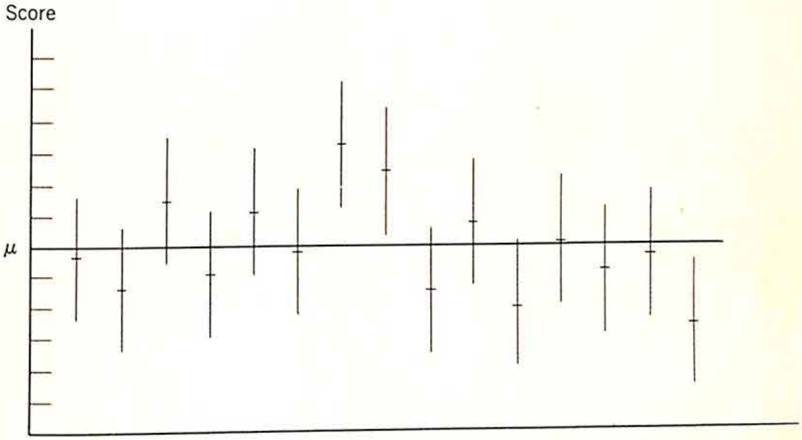


FIG. 7.3. Illustration of confidence intervals for population mean.

The relation between the confidence interval and the procedure in testing a hypothesis should be noted. For instance, in an earlier section of this chapter we discussed a two-sided test of the hypothesis $H : \mu = 73$. The sample mean, $\bar{X}$, was 66, the standard deviation was taken to be 12, n was 16, and $\sigma_{\bar{X}}$ was 3. From the observed mean, $\bar{X} = 66$, we could compute 95 percent confidence limits: $66.00 - 5.88 = 60.12$ and $66.00 + 5.88 = 71.88$. The interval 60.12–71.88 covers the values of μ which would be "accepted" at the 5 percent level of significance. The value 73 is *not* contained in the 95 percent confidence interval. Therefore we reject the hypothesis that $\mu = 73$ at the 5 percent level.

7.9 ESTIMATING VARIANCE FROM A SAMPLE

Up to this point in our discussion of the application of the sampling distribution of the mean, we have depended upon *known* values of σ. Most frequently the investigator does not know the variance of the population, and must estimate it from the sample itself. This complicates the problem of statistical inference and the calculation of confidence intervals regarding population means. With large samples, from populations for which $\bar{X}$ is normally distributed, the sample value of $\Sigma x^2/n$ may be taken as an estimate of σ^2, and the foregoing procedure is appropriate. However, we will now develop a better approach.

We have concentrated our attention on the sampling distribution and some of the estimation problems relating to the mean. Nothing has been said about estimating the variance or the sampling distribution of the variance. In Chapter 10 we will discuss the sampling distribution of the variance. For the present we require only some knowledge concerning the estimation of the variance. If we define the statistic

$$s^2 = \frac{\Sigma(X - \bar{X})^2}{n - 1} \tag{7.3}$$

it can be shown that

$$E(s^2) = \sigma^2$$

where, as previously, σ^2 is the population variance. If the expectation of an estimator is the parameter itself, the estimator is said to be unbiased. Thus, s^2 is an unbiased estimate of σ^2.

We will henceforth use s^2 as defined in equation 7.3 in computing variances from samples, and will refer to it as the sample variance. In Chapter 4 we used the formula $\Sigma(X - \bar{X})^2/n$. In that chapter we were not considering the sampling problem. When n is large, of course, it makes little difference whether the denominator is n or $n - 1$.

There are two systems of notation in current use. The symbol s^2 is used by some authors for $\Sigma(X - \bar{X})^2/n$ and by others as defined in equation 7.3. The student should learn to distinguish between the two notations. If the variance is computed as

$$s_b^2 = \Sigma x^2/n$$

then an unbiased estimate of σ^2 would be

$$\begin{aligned} s^2 &= \frac{\Sigma x^2}{n - 1} = \frac{\Sigma x^2}{n} \cdot \frac{n}{n - 1} \\ &= \frac{n}{n - 1} s_b^2 \end{aligned}$$

with s^2 defined as in equation 7.3. Both $\Sigma x^2/n$ and $\Sigma x^2/(n - 1)$ are estimates of the population variance, but only the latter is *unbiased.*

Now if we let $s_{\bar{X}}^2$ represent an estimate of $\sigma_{\bar{X}}^2$, from equation 7.1.

$$\begin{aligned} s_{\bar{X}}^2 &= s^2/n \\ &= \frac{\Sigma(X - \bar{X})^2}{n(n - 1)} \end{aligned} \tag{7.4}$$

and

$$s_{\bar{X}} = s/\sqrt{n}$$

A verification of equation 7.3 occurs in Table 7.1. Entered in the fourth column of that table are the sums of squares for each of the 16 possible samples. Since in the experiment $n = 2$, the denominator of s^2 is 1. Therefore, the entries in column 4 are also the sample estimates of the population variance which we have already shown to be $\sigma^2 = 1.25$. The average of these 16 values of s^2 is seen to be $20/16 = 1.25$. This demonstrates (though it does not prove) that s^2 is an unbiased estimate of σ^2.

We observe furthermore that the estimates of σ^2 vary from 0 to 4.50, even though they average 1.25. If now we compute from each sample an estimate of $\sigma^2_{\bar{X}}$, from equation 7.4, we would have $.00/2 = .00$; $.50/2 = .25$; $2.00/2 = 1.00$; etc. The total of all 16 such values of $s^2_{\bar{X}}$ would be $20/2 = 10$. The average would be $10/16 = .625$, our computed *true* value of $\sigma^2_{\bar{X}}$. We thus verify that $s^2_{\bar{X}}$ is indeed an *unbiased* estimate of the variance of sample means.

7.10 SIGNIFICANCE OF A MEAN—σ^2 ESTIMATED

In the sections on confidence intervals and the testing of statistical hypotheses, we made use of the knowledge that in a normal (or near normal) distribution the statistic $(\bar{X} - \mu)/\sigma_{\bar{X}}$ is itself normally distributed (or approximately so). On the strength of the previous section we might be tempted to use $(\bar{X} - \mu)/s_{\bar{X}}$ as if it were normally distributed when our only information is from the sample itself. This is precisely what is done frequently in practice. However, this procedure is adequate only when samples are very large.

In small samples, s will not, except by accident, be the same as σ. It is a sample *statistic*, just as is $\bar{X}$, and subject to sampling variation, just as $\bar{X}$ is subject to variation from sample to sample. In the ratio, $z = (\bar{X} - \mu)/\sigma_{\bar{X}}$, the divisor is a known constant ($\sigma_{\bar{X}}$ being the true or known standard error of the mean, a parameter). On the other hand, the ratio

$$t = (\bar{X} - \mu)/s_{\bar{X}} \tag{7.5}$$

contains a sample value in both numerator and denominator. We use the normal distribution to find probabilities when sampling from normal distributions (or when samples are very large)—*and when σ is known.* Technically then, *and only then*, are we justified in using the normal distribution as the distribution of z. The statistic t as defined by equation 7.5 has a known sampling distribution when the variate, X, is itself normally distributed. The distribution of t is not the same as the normal distribution, although it approximates the normal curve for large n. It is sometimes called Student's distribution after the pseudonym used by W. S. Gosset, who published a famous paper on the subject in 1908.

In many respects the Student's, or *t* distribution, is similar to the normal distribution. It is symmetrical, and it is bell shaped. The chief difference is that it is really a whole family of distributions—one for each value of the number of *degrees of freedom*, defined as the divisor of $\Sigma(X - \bar{X})^2$ which yields an unbiased estimate of σ^2, usually $(n - 1)$.

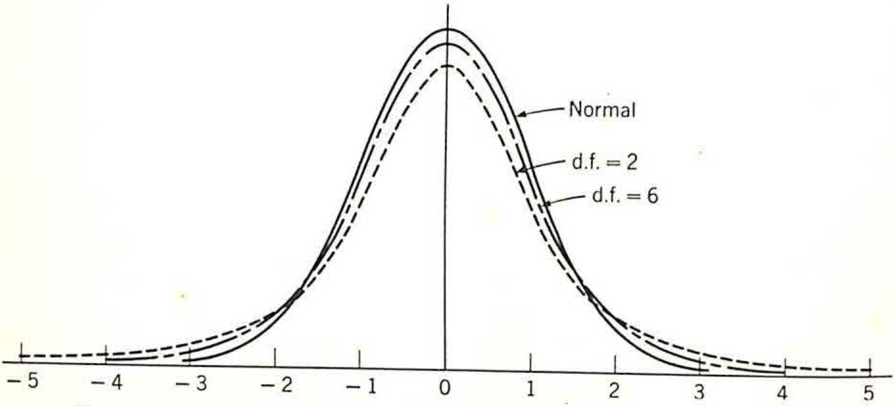


FIG. 7.4. Comparison of normal and *t* distributions for 2 and 6 degrees of freedom.

Figure 7.4 is a diagram comparing the normal probability distribution and the distribution of *t* for $(n - 1) = 2$, that is, 2 degrees of freedom, and for $(n - 1) = 6$ degrees of freedom.

It may be seen that *t* "spreads out" more than the normal curve, and the fewer the degrees of freedom the more the spread. This characteristic of the *t* distribution is also shown in Table 7.5, which gives for various

TABLE 7.5

VALUES OF *t* WHICH WILL BE EXCEEDED NUMERICALLY 5 PERCENT OF THE TIME

d.f.	t*
∞	± 1.96
30	± 2.04
20	± 2.09
10	± 2.23
5	± 2.57
4	± 2.78
3	± 3.18
2	± 4.30
1	±12.71

* Values of *t* which cut off 2.5 percent of area under the probability distribution of *t* on each tail, $t_{.025}$ and $t_{.975}$.

degrees of freedom the values of $\pm t$ which contain the middle 95 percent of the t values. Note, for instance, that when d.f. $= 2$, that is, usually when $n = 3$, the middle 95 percent of t values ranges as far as ± 4.30.

It will be recognized in Table 7.5 that the value of $t_{.975}$ for infinity is 1.96, the same as the value of $z_{.975}$, which we used in testing hypotheses at the 5 percent level of significance and in establishing 95 percent confidence intervals.

For small samples we see that the value of $t_{.975}$ at the 5 percent level is greater. The same is true of other levels. Hence, the probability of a specified large deviation from the mean (zero) increases as the degrees of freedom decrease. Since in small samples the t distribution contains more area in the tails, larger values of t are required to indicate significance at a given level than in large samples.

Extensive tables have been published, like tables for the normal curve, giving integrals and the cumulative frequency function $F(t)$ for various degrees of freedom. In the table of Appendix E are shown only those values of t which are most frequently of interest in testing statistical hypotheses and in establishing confidence intervals. Appendix E should be used to verify values in Table 7.5. Since t is distributed symmetrically, $t_{.025} = -t_{.975}$ just as $z_{.025} = -z_{.975}$. Similarly, $t_{.05} = -t_{.95}$ just as $z_{.05} = -z_{.95}$. It will be recognized that the values in Table 7.5 are those which cut off 2.5 percent of the area from each tail. The values are those which may, therefore, be used in a *two-tailed test* at the 5 percent *level of significance*, or in establishing 95 percent confidence intervals.

As with the normal curve, there is (for each number of degrees of freedom) a frequency function $f(t)$ and a cumulative distribution function $F(t)$. In symbols, we write, for any given degrees of freedom,

$$P(|t| > t_0) = 2\int_{t_0}^{\infty} f(t)\,dt = 2[1 - F(t)]$$

where t_0 is some specified value of t.

Similarly,

$$P(t > t_0) = \int_{t_0}^{\infty} f(t)\,dt = 1 - F(t)$$

In the first case, area is included in both tails; in the latter in only one tail. Tables of Student's distribution are given in both forms, and it is thus important to be able to recognize the distinction. Probabilities in the former case must be halved for one-tailed tests. The positive values in Table 7.5, for example, are those for which $P(t > t_0) = .025$, not .050.

The similarity of Student's distribution and the normal distribution should permit an easy transfer of our experience in sampling application of the latter to uses of the former.

A confidence interval for the population mean may be established on the basis of the inequality

$$(\bar{X} - t_0 s_{\bar{X}}) < \mu < (\bar{X} + t_0 s_{\bar{X}}) \tag{7.6}$$

where t_0 is the value of t corresponding to the desired confidence coefficient. For example, the following are scores for speed of oral reading measured in words per second on a sample passage, for a sample of seven ninth-grade pupils whose marks and achievement test scores were found to be below average and below expected performance as determined by intelligence tests: 3.7, 3.1, 4.4, 4.8, 4.6, 3.6, 4.5. We find $\bar{X} = 4.10$, and compute $\Sigma(X - \bar{X})^2$ from $\Sigma X = 28.7$ and $\Sigma X^2 = 120.07$ as follows: $\Sigma X^2 - (\Sigma X)^2/n = 120.07 - 823.69/7 = 2.40$

Then, from equation 7.4,

$$s_{\bar{X}}^2 = 2.40/42 = .0571$$

and, taking the square root, we obtain

$$s_{\bar{X}} = .239$$

Assuming that we are interested in a 99 percent confidence interval, we find $t_{.995} = 3.71$ for d.f. $= 6$. From equation 7.6 we find the 99 percent confidence limits, $\bar{X} \pm t_0 s_{\bar{X}}$ to be $4.10 \pm (3.71)(.239)$, or 3.21 and 4.99. We would *accept* any hypothetical μ_0 between these limits at the 1 percent level and *reject* any μ_0 outside these limits. For instance, $H : \mu = 5.2$ would be rejected.

We could test this hypothesis directly by means of $t = (\bar{X} - \mu)/s_{\bar{X}}$, rejecting if $t < t_{.005}$ or $t > t_{.995}$ for 6 d.f. In this particular case, according to this hypothesis, from equation 7.5 we could compute t as follows:

$$t = \frac{4.10 - 5.20}{.239} = -1.10/.239 = -4.60$$

which is clearly in the rejection region specified at the 1 percent level.

It should not be overlooked that the t function tabled in Appendix E applies only to *random* sampling, and strictly only to sampling from *normal* parent populations. Its use in establishing confidence intervals or in testing statistical hypotheses is invalid if applied to "purposive samples," "samples of convenience," or chunks, or if the assumption of normality in in the parent population is not reasonably tenable.

7.11 SAMPLING WITHOUT REPLACEMENT—FINITE UNIVERSE

Most of the discussion in this chapter has dealt with formulas and theory which are applicable only to sampling *with replacement* or, as we have indicated, from such a large universe that the distinction is not important. In actual practice, we often sample *without* replacement from finite populations.

It can be shown that $\frac{N-1}{N}s^2$ is an unbiased estimate of σ^2 when sampling *without* replacement, where N is the size of the population. That is, an unbiased estimate of the population variance is

$$\hat{\sigma}^2 = \frac{N-1}{N}s^2 \tag{7.7}$$

when sampling without replacement. In very large universes, of course, the factor $(N-1)/N$ is negligible. In a population of 500, for instance, this factor does not amount to much. In a population of 30, however, it is .967.

Also the standard error of the mean, $\sigma_{\bar{X}}$, in a finite population (when sampling *without* replacement) differs from one from an infinite population by a factor $(N-n)/(N-1)$, where N is the population size and n the sample size. Hence,

$$\sigma^2_{\bar{X}} = \left(\frac{N-n}{N-1}\right)\frac{\sigma^2}{n} \tag{7.8}$$

and

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \tag{7.9}$$

are formulas which we would properly have used in Section 7.3 if we had sampled *without* replacement. However, with $N = 711$ and $n = 10$ or $n = 20$, the ratio of N to n is so large that there would be little difference between equations 7.2 and 7.9. The *finite multiplier*, as the ratio $(N-n)/(N-1)$ is often called, influences the value of $\sigma_{\bar{X}}$ only according to its square root. A $\sigma_{\bar{X}}$ estimated by sampling from a finite population of 500 with $n = 25$ would be 97.5 percent of $\sigma_{\bar{X}}$ from an infinite population. The factor would amount to 87 percent, where $N = 100$ and $n = 25$. The consequence of not using the finite multiplier in equation 7.9 would result in a value too large by 15 percent. The example of Table 7.1 may be converted to sampling from a finite population *without* replacement simply by eliminating the four starred samples. It is left to the student to compute new totals for the table and to verify that under these conditions $\bar{\bar{X}} = 2.50$; that from equation 7.7 the "expected" value or average value of $\frac{N-1}{N}s^2$ is 1.25, the population variance; and that from equation 7.8 the variance of the means is

$$\sigma^2_{\bar{X}} = \left(\frac{4-2}{4-1}\right)\frac{1.25}{2} = .4167$$

When s^2 is used to estimate σ^2 for a finite population, it should be multiplied by $(N-1)/N$ because of equation 7.7. Hence, from equation 7.8

$$s^2_{\bar{X}} = \frac{s^2}{n}\left(\frac{N-1}{N}\right)\left(\frac{N-n}{N-1}\right) = \left(\frac{N-n}{Nn}\right)s^2 = \left(\frac{N-n}{N}\right)\frac{s^2}{n} \qquad (7.10)$$

an unbiased estimate of the variance of the mean.

Suppose a sample of enrollments of 20 isolated one-teacher schools obtained from 8 of the 20 schools is as follows: 5, 16, 21, 12, 19, 10, 17, 18. We compute $\Sigma X^2 = 1{,}940$; and $\Sigma X = 118$. The mean, $\bar{X}$, is $118/8 = 14.75$. The sum of squares is $\Sigma x^2 = \Sigma X^2 - (\Sigma X)^2/n = 1940 - 1{,}740.50 = 199.50$. The sample variance is

$$s^2 = 199.50/7 = 28.5$$

From equation 7.10

$$s^2_{\bar{X}} = \left(\frac{12}{160}\right) 28.5 = (.075)(28.5) = 2.14$$

and

$$s_{\bar{X}} = 1.46$$

The coefficient of variation, C.V., is $s_{\bar{X}}/\bar{X} = 1.46/14.75 = .099$. The error variance of the mean is thus estimated to be about 10 percent of the mean. Approximate 95 percent confidence limits, $\bar{X} \pm 2s_{\bar{X}}$, would be 14.75 ± 2.92, that is, 11.8 and 17.7. When n and $N - n$ are large, $\bar{X}$ is approximately normally distributed.

7.12 THE DISTRIBUTION OF AN ESTIMATED TOTAL

Frequently the objective is an estimate of an "aggregate" or total, $\sum_{i=1}^{N} X_i$. One estimate of the total enrollment of the twenty schools would be the mean multiplied by the number of schools in the population, $N\bar{X} = (20)(14.75) = 295$. This is an unbiased estimate. The sampling variance of this estimate is closely related to that of $\bar{X}$.

We know that if we multiply each value in a distribution by a constant a; the variance of the resulting distribution is $a^2\sigma^2$ and $a\sigma$ is the standard deviation. If we were to multiply each of all possible sample means, $\bar{X}_i$, by N, the finite population size, the variance of the resulting distribution would be $N^2\sigma^2_{\bar{X}}$. This would be the variance of the statistic $N\bar{X} = \frac{N}{n}\sum_{i=1}^{n} X_i$, an estimate of the grand total, or "aggregate," of the population of N items.

Where σ^2 is known, the variance of the estimated total, $N\bar{X}$, of N items is thus

$$\sigma_S^2 = N^2\sigma^2/n \tag{7.11}$$

in sampling an *infinite* population, and

$$\sigma_S^2 = \left(\frac{N-n}{N-1}\right)\frac{N^2\sigma^2}{n} \tag{7.12}$$

in sampling a *finite* population.

Where σ^2 is not known, and we must estimate it from a sample, the estimated variance of $N\bar{X}$ is

$$s_S^2 = N^2s^2/n \tag{7.13}$$

in sampling an *infinite* population, and

$$s_S^2 = \frac{N^2s^2}{n}\frac{(N-n)}{N} = \frac{N(N-n)}{n}s^2 \tag{7.14}$$

in a *finite* population.

The estimated variance of the distribution of sample estimates, $N\bar{X}$, for the twenty schools would be, from equation 7.14,

$$s_S^2 = \frac{20(12)}{8}28.5 = 855$$

whence,

$$s_S = \sqrt{855} = 29.24$$

The coefficient of variation of the aggregate is C.V. $= (Ns_{\bar{X}}/N\bar{X}) = (s_{\bar{X}}/\bar{X})$, the same as the coefficient of variation for the mean. In this case it is $29.24/295 = .099$.

7.13 THE STANDARD ERROR OF A PROPORTION OR PERCENT

In Section 6.5 we saw that the normal distribution under suitable conditions is an acceptable approximation to the binomial distribution. When the proportion p is known, and not too near 0 or 1, and the size of the sample n is not too small, we may apply directly the methods of Sections 7.4 and 7.8 to sample data from binomial distributions. For this purpose we compute the population mean $\mu = np$ from equation 5.8 and the variance $\sigma^2 = npq$ from equation 5.9 and proceed by means of the normal distribution.

Under these conditions we use the normal distribution to establish limits between which we would expect any given proportion of X values. These limits are given by

$$np \pm z_0 \sqrt{npq} \tag{7.15}$$

where z_0 is the deviate of the unit normal curve appropriate to the desired level of confidence.

On the other hand, we could find the *standard deviate*, z_1, which tells us the amount by which an observed sample number, X, possessing a given characteristic, varies from the population mean np thus:

$$z_1 = \frac{X - \mu}{\sigma} = \frac{X - np}{\sqrt{npq}} \tag{7.16}$$

For example, if $n = 600$ and $p = .6$, then $np = 360$ and $npq = \sqrt{144} = 12$. If we choose $z_0 = 2.00$, then $np \pm 2\sqrt{npq} = 360 \pm 24$ and $P[(336) < X < (384)]$ is approximately .95. An observed number of "successes" of $X = 300$ is outside such limits and would be judged an unusually rare chance occurrence in assessing the significance of such a sample. Or in testing a hypothesis $H: \mu = 360$ at, for instance, the 1 percent level, we would compute, as in Section 7.5, the statistic, $z_1 = \dfrac{X - np}{\sqrt{npq}} = \dfrac{300 - 360}{12} = -5.0$. Since this is outside ± 2.58 we reject the hypothesis.

The standard deviation of the proportion[1] p is $1/n$ times the standard deviation of $\mu = np$. That is

$$\sigma_p = \frac{1}{n}\sqrt{npq} = \sqrt{pq/n} \tag{7.17}$$

In the above example the observed proportion, $p_1 = X/n = 300/600 = .5$. We could compute $z_1 = \dfrac{(X/n) - p}{\sqrt{pq/n}} = -5.0$, the same result as above. In other words, our original test was the equivalent of testing the hypothesis that there is no difference between the population *proportion* $p = .6$, and the sample *proportion*, $p_1 = .5$.

Now suppose that we have no *a priori* probability; p is unknown though we assume X to be binomially distributed. For instance, 240 of a sample of 400 ninth-grade students "pass" a test item. The observed $p' = X/n = 240/400 = .60$. One method is to take p' directly as an

[1] The procedure is identical if p is reported as percent; results will then be also in percent.

estimate of the parameter p, and compute an estimate of σ_p^2. An unbiased estimate of σ_p^2 is

$$s_p^2 = \frac{p'q'}{n-1} \tag{7.18}$$

From this we may compute the standard error

$$s_p = \sqrt{\frac{p'q'}{n-1}} \tag{7.19}$$

The difference between the use of equations 7.17 and 7.19 is not great when n is large, say, 30 or more. The former is frequently used in any event although the latter is more suitable for the same reason that we used $(n-1)$ as a divisor in estimating variance in equation 7.3.

In our example

$$s_p = \sqrt{p'q'/(n-1)} = \sqrt{.24/399} = .0245$$

Approximate 95 percent confidence limits would be

$$p' \pm 1.96\sqrt{\frac{p'q'}{n-1}} \tag{7.20}$$

In this case the limits would be $.60 \pm (1.96)(.0245)$, or .552 and .648.

When n is large and X/n near .5, this is a satisfactory approximation. A better approximation to confidence intervals for a proportion is based upon the equation

$$\frac{X/n - p}{\sqrt{pq/n}} = \pm z_0 \tag{7.21}$$

You will recognize equation 7.21 as the limits of the confidence interval with z_0 corresponding to whatever confidence level we choose. (For 99 percent confidence level z_0 would be 2.58; 95 percent would be 1.96; etc.). Squaring equation 7.21 and solving for the only unknown, p, $(q = 1 - p)$, we obtain

$$p = \frac{(X + z_0^2/2) \pm z_0\sqrt{X(n-X)/n + z_0^2/4}}{n + z_0^2} \tag{7.22}$$

In a sample for which $X = 240$ and $n = 400$ the results would be .551 and .647 differing little from results obtained earlier by means of equation 7.20. When n is smaller, or p farther from .5, equation 7.22 provides a more accurate approximation.

Considerable computation may be avoided by means of published charts of "confidence regions" for given confidence levels of proportions.[1]

[1] For reproductions of the original Clopper-Pearson charts see Dixon and Massey.[2]

As the discussion in Section 6.5 shows, there are limits to the goodness with which the normal distribution approximates the binomial. As a general rule caution should be exercised when np or the observed X is small, say, less than 5. In practice it may be worthwhile to check such approximate tests of significance when observed values just reach or just fail to reach the critical region, by an exact calculation using the binomial expansion itself. It should not be forgotten that there are tables of the binomial for small values of n to which reference was made in Chapter 5, reference 3.

It should be clear from the definition of measurement in Chapter 2 and the discussion of Section 6.5 that there should be a *correction for continuity* when applying the normal curve to a binomial distribution. For instance, in the binomial for which $n = 12$ and $p = .65$, we might be interested in $P(5 < X < 10)$. By the methods of Chapter 5 we would define the interval limits as 5.5 and 9.5 and by means of equation 7.16 find the corresponding variates z_1 and z_2 as follows:

$$z_1 = \frac{5.5 - (12)(.65)}{\sqrt{(12)(.65)(.35)}} = -1.39$$

and

$$z_2 = \frac{9.5 - (12)(.65)}{\sqrt{(12)(.65)(.35)}} = +1.03$$

From the table of integrals we find the two areas on each side of the mean to be .4177 and .3485. Therefore, $P(5 < X < 10) = .7662$. The correction for continuity is usually disregarded when X and n are large.

We can write the approximate probability, $P(X_1 \leq X \leq X_2)$, of X occurring between *and including* the two values X_1 and X_2, where $(X_1 < X_2)$, as the area under the unit normal curve between z_1 and z_2, where

$$z_1 = \frac{(X_1 - \frac{1}{2}) - np}{\sqrt{npq}}, \quad z_2 = \frac{(X_2 + \frac{1}{2}) - np}{\sqrt{npq}} \tag{7.23}$$

Where sampling is *without replacement* from finite populations the *finite multiplier* must be used as in equations 7.8 and 7.9. Hence, where p is the *population parameter*, and other notation is as before, the variance of X in a binomial distribution is

$$\sigma_x^2 = npq\left(\frac{N - n}{N - 1}\right) \tag{7.24}$$

and the variance[1] of p is

$$\sigma_p^2 = \frac{pq}{n}\left(\frac{N - n}{N - 1}\right) \tag{7.25}$$

[1] When p is estimated from a sample, an unbiased estimate of σ_p^2 is $s_p^2 = \dfrac{(N - n)}{(n - 1)}\dfrac{p'q'}{N}$. This differs little from equation 7.25.

With a correction for continuity, the confidence limits for a percent or proportion from a finite population sample may be approximated as in equation 7.20 by

$$p \pm \left\{z_0 \sqrt{\frac{N-n}{N-1}} \sqrt{\frac{pq}{n}} + \frac{1}{2n}\right\} \qquad (7.26)$$

We note that the discussion in this section has not included reference to Student's distribution. The *t* distribution is appropriate when the population variance must be estimated from the data. However, in testing the deviation of an observed proportion from a hypothetical proportion, the variance of sample values around the hypothetical proportion is fixed. Hence, the *t* distribution is not applicable to the binomial problems which have been discussed in this section.

EXERCISES

1. Define:

Sampling distribution
Standard error
Variance of a mean
Statistical hypothesis
Test of a statistical hypothesis
Critical region
Level of significance
Power of a test
Confidence interval
Confidence limits
Level of confidence
Confidence coefficient
Unbiased estimate
Correction for continuity
Degrees of freedom
t distribution
Student's distribution
Finite multiplier
Coefficient of variation

2. Treat the California Test scores of Appendix A as an infinite population. This can be done by drawing samples from them *with replacement*. The experiment will be to verify Theorems *a*, *b*, *c*, and *d* from samples of sizes 10 and 20. As a class exercise, each student should be allotted a quota of an even number of samples of size 10 to be drawn from this distribution so that there will be a total of at least 200 such samples for the entire class. Assign serial numbers to the 168 scores, beginning with 001 and ending with 168. By means of the table of random numbers in Appendix B find separate sets of 10 random "three-digit" numbers to determine which ten scores of the 168 constitute a sample. For *each* sample of 10 compute:

(*a*) The mean, $\bar{X}$.

(*b*) The sample variance from formula 7.3.

(*c*) The estimated variance of the sample mean from formula 7.4.

(*d*) The 70 percent confidence interval, using formula 7.6 with $t_0 = 1.10$.

Combine consecutive samples of 10 in pairs to obtain means of samples of 20 (at least 100 such samples). The work may be simplified by combining the sums, ΣX, of the two samples of 10. For each sample of 20 compute the mean.

Prior to drawing samples, work up a standard computation sheet, showing by column and row the values to be entered for each sample. For instance, each sample of 10 might be assigned a line (or row) on the work sheet. Columns 1 to 10 could be individual scores. Additional columns would contain ΣX, ΣX^2, $\bar{X}$, $\bar{X}^2$, Σx^2, s^2, $s_{\bar{x}}^2$, $s_{\bar{x}}$, $1.10 s_{\bar{x}}$, $\bar{X} - 1.10 s_{\bar{x}}$, $\bar{X} + 1.10 s_{\bar{x}}$.

Combine all samples of 10 for the class to show: (*a*) The mean of sample means. (*b*) The variance of the sample means. (*c*) The number and proportion of confidence intervals which contain the population mean, 70.08. (*d*) A frequency distribution of the means.

Compare results with theoretical values, given the population variance, $\sigma^2 = 174.09$. Plot the cumulative frequency distribution of the sample means on normal probability paper.

Combine all samples of 20 for the class to show: (*a*) The variance of the sample means. (*b*) A frequency distribution of the means. Compare results with theoretical values and explain. Plot the cumulative-frequency distribution on probability paper.

3. Among 100 digits selected at random from a table in which odd and even digits appear equally often, it was found that 60 even digits were drawn. What is the probability that there would be a discrepancy as great or greater than this?

4. By means of the normal approximation to the binomial what would be the probability of answering correctly, *by guessing alone*, as many as 17 of 50 four-choice items on a test?

5. A *lowest nonchance* score is to be established as one which exceeds the average chance score by twice the standard error of the mean chance score. What is the lowest nonchance score for a 25-item test of five-choice items?

6. A roulette wheel at a county fair has 36 numbers. It is observed that 16 times in 1,000 trials the wheel has stopped on 24. Test the hypothesis that the wheel is not "fixed."

7. In a study of reading, measurements were obtained on the number of fixations per line of print for 14 first-grade subjects. Given $\Sigma X = 260$, and $\Sigma X^2 = 4{,}881$, find 99 percent confidence limits for the mean.

8. Scores on an academic aptitude test were administered to a sample of 17 of 48 seniors in a high school. Given $\Sigma X = 1{,}246$ and $\Sigma X^2 = 91{,}931$, find the sample mean, the standard error of the mean, and 95 percent confidence limits for the mean. Test the hypothesis $H : \mu = 71.00$.

9. A sample of 35 of 87 high-school home economics departments in a state were found to have an aggregate (ΣX) of 273 sewing machines. What is an unbiased estimate of the total number of sewing machines in all 87 departments? If $\Sigma X^2 = 2{,}756$, find the standard error of the estimated total. What is the coefficient of variation? Confidence limits at the 95 percent level?

10. Assume the 13 female noncollege students in high school C of Appendix G to be a random sample of a very large population.

(*a*) What would be the unbiased estimate of the population CTMM score variance?

(*b*) The standard error of the sample mean?

(*c*) What are the 95 percent confidence limits for the mean?

(*d*) Supposing the population to be defined as a finite population with $N = 100$, what effect would this have on your answers to *a*, *b*, and *c*?

11. Suppose that the information available to you from Appendix G is only that from the 45 CTMM scores from high school A. Suppose further that this is defined as a random sample from a very large population. At the .01 level, test the hypothesis that the population mean is 80 or more.

12. What is the effect upon $\sigma_{\bar{x}}$ of doubling sample size? Of making the sample size five times as large? Of increasing sample size *r* times its size?

13. A sample of eight observations is taken from a population of known variance. A test of a hypothesis concerning the population mean is made by computing the statistic, $z = (\bar{X} - \mu)/\sigma_{\bar{x}}$, and entering a t table for the 5 percent value of t for 7 d.f. Is the correct value of α

(*a*) .05? (*b*) less than .05? (*c*) greater than .05?

14. Plot the distributions of Table 7.3 on normal probability paper. Would you expect a large number of samples of size 50 to be more like the normal or less like the normal?

15. Test the hypothesis at the .01 level that the 35 CTMM scores for high school B of Appendix G constitute a random sample from a population with $\sigma^2 = 196$ and $\bar{X} = 70$.

16. From the manual of a test it is found that the norm for ninth-grade students is 83 and the standard deviation 10. How many ninth-grade students should be included in a sample so that the mean may be estimated within 5 percent with 95 percent confidence?

17. An opinion poll is to be undertaken before a school election to estimate the percent of voters who will be in favor of a proposal. It is expected that the election will be close so that the proportion, p, of voters favoring the proposal will be near .50. Estimate the size of sample required for a random sample of voters so that the sample proportion, p', will be within .02 of the true value of p with 95 percent confidence.

18. A group of 36 fifth-grade pupils are to be given instruction in reading by what is claimed to be an improved method. A criterion test of achievement is to be used for which the fifth-grade norm is known to be a score of 75 and for which the population variance is estimated to be 144. Assuming the variance of scores to be 144, and assuming that interest lies only in a one-sided test:

(*a*) Write in symbols and in words the hypothesis to be tested.

(*b*) What is the rejection region for testing this hypothesis at the .05 level?

(*c*) What is the power of this test at the 5 percent level in rejecting the hypothesis if the true value of μ is 74? If the true value of μ is 78?

19. In testing a hypothesis, $H : \mu = \mu_0$, concerning a mean with σ known, a critical region is chosen so that the hypothesis will be rejected if $z < -2.58$ or if $z > +2.58$.

(*a*) Sketch a normal curve showing the region of rejection.

(*b*) What is the Type I error if the hypothesis is true?

(*c*) What is the Type I error if the hypothesis is not true?

(*d*) What is the value of α if the *true* mean is $\mu_0 + 2\sigma_{\bar{x}}$? If the *true* mean is $\mu_0 - 2\sigma_{\bar{x}}$?

(*e*) What is the Type II error if the *true* mean is $\mu_0 + 2\sigma_{\bar{x}}$? If the *true* mean is $\mu_0 - 2\sigma_{\bar{x}}$? Add necessary freehand lines to your sketch for *a* above to show this.

(*f*) What is the power of the test if the *true* value of the mean is $\mu_0 \pm 2\sigma_{\bar{x}}$? Identify that area in your sketch which represents $1 - \beta$.

20. A population is known or is estimated to have a standard deviation of 20. A hypothesis is to be tested at the 5 percent level that $\mu = 60$ (a two-sided test). Since the standard deviation of the population is presumed to be known, the rejection region is in terms of the normal deviate ($z < z_{.025}$ and $z > z_{.975}$). The results are to be used in an experiment for which it is important to have some assurance that μ is not less than 50 or greater than 70.

(*a*) Suppose that a sample size of $n = 4$ is to be used. What is the probability of rejecting the hypothesis if it is indeed false and that in fact $\mu = 50$? If in fact $\mu = 70$? In either case, what is the power of the test?

(*b*) Is the power of the test at least as great as this if μ is in fact less than 50? Greater than 70? Explain.

(*c*) What is the power of the test in guarding against the alternative hypothesis that μ is less than 50 or greater than 70 when the sample size is $n = 9$? When $n = 16$? When $n = 25$?

(*d*) What effect does increasing sample size have upon the power of the test?

21. Several schools are to be selected for an experiment. The conditions of the experiment require minimizing the risk of including student bodies with below average I.Q. as measured by a special intelligence test. Assume the standard deviation of the I.Q. distribution to be 16.

(*a*) In terms of $z = (\bar{X} - \mu)/\sigma_{\bar{x}}$ what is the rejection region if it is desired to be at least 90 percent sure of accepting a school if its mean I.Q. is 105 or more?

(*b*) In terms of z what is the rejection region if it is desired to reject a school with a probability of at least .90 if its mean I.Q. is 100 or less?

(*c*) Suppose that it is desired both to be at least 90 percent sure of accepting if μ is 105 or more and to be at least 90 percent sure of rejecting if μ is 100 or less. Draw a diagram on a scale marked off in intervals of z, sketching in normal curves and identifying the critical region for each of the two requirements.

(*d*) On the diagram of *c* show the number of units of z from 100 to the critical value and the number of units of z from 105 to the critical value. How many units of z must cover the distance from 100 to 105?

(*e*) What, therefore, must be the value of $\sigma_{\bar{x}}$?

(*f*) Equating the result of (*e*) to $\sigma/\sqrt{n}$ and substituting the assumed value $\sigma = 16$, find the number of cases necessary in order to meet both of the conditions specified in *c*.

22. An experiment is designed to compare tire mileages of two brands of tires used on school buses. Two tires of each brand are used on each school bus. The experiment is repeated on 20 school buses. Six of the 20 experiments came out in favor of brand A and 14 in favor of brand B. At the .05 level, test the hypothesis that in a population of such experiments there would be the same number of results favoring one brand as there would be for the other.

REFERENCES

1. Cornell, Francis G., "Sample Surveys in Education," *Review of Educational Research*, 20: 227-248, September 1955.
2. Deming, William E., *Some Theory of Sampling*, New York, John Wiley and Sons, 1950, Chapter 10.
3. Dixon, Wilfrid J., and Frank J. Massey, *Introduction to Statistical Analysis*, New York, McGraw-Hill Book Co., 1951, Chapters 7, 9, and 14.
4. Freund, John E., *Modern Elementary Statistics*, New York, Prentice-Hall, 1952, Chapters 8, 9, and 10.
5. Hansen, Morris H., William N. Hurwitz, and William G. Madow, *Sample Survey Methods and Theory*, New York, John Wiley and Sons, 1953, Vol. 1.
6. Johnson, Palmer O., *Statistical Methods in Research*, New York, Prentice-Hall, 1949, pp. 33-37.
7. Wilks, Samuel S., *Elementary Statistical Analysis*, Princeton, N. J., Princeton University Press, 1949, Chapters 10 and 11.

CHAPTER 8

Correlation and Regression

The purpose of this chapter is to present basic concepts for measuring the relationship of one variable to another. Earlier chapters in this book have dealt with a single variable. The most common devices for describing the relationship of pairs of measures of two different variables are the *correlation coefficient* and the *regression coefficient*.

The correlation coefficient is a much-used tool of the educational researcher and measurement specialist. In various forms it has been the foundation of educational measurement. It has also served yeoman duty in analytical research. Newer methods of experimentation involving some of the theory of Chapter 7 and later chapters are now vying with correlation methods for honors in the research arena. Developments in statistical theory continue to produce systems of analysis which are reducing the emphasis on correlation methods. Nevertheless, the correlation coefficient and its companion the regression coefficient will continue to have a prominent role among the techniques of educational research.

It is important to know the relationship of many things in education—the relationship between a classification or an aptitude test and a measure of proficiency or achievement; between the cost of a service and characteristics of a service; between qualities of teachers and measures of the performance of teachers in their jobs; between amounts and kinds of teaching or learning experience and the amount of learning. There are many such realities of how education is organized and how it takes place about which little is known.

The experimenter may control variable *a* to study its effects on variable *b*. Unfortunately, many human variables in the social milieu of the classroom, the school system, or the community escape the manipulation of the laboratory. The relationships of these, however, may often be charted and measured by means of correlation methods.

8.1 THE MEANING OF COVARIANCE

There are many circumstances which involve more than one measurement of a set of subjects, observations, objects or operations. A typical pupil record contains many variables, such as age, height, weight, years in school, grades or marks in various subjects, mental age, IQ, scores from tests of achievement, aptitude, interest, personality adjustment, and such. For the time being we consider the relationship only of pairs of such measures. In a later chapter we consider methods of treating sets of three or more measures from a common population or sample.

For many purposes pairs of measures are recorded for a large sample or an entire population. It is not unusual in standardizing tests to deal with pairs of scores, for example, scores on two parts of a test, for 5,000 subjects. It is frequently necessary to work with pairs of scores, on the other hand, from a small sample. To introduce principles we shall use a small sample.

The data of Table 8.1 are adapted from an experiment in pilot training conducted in a flying school. There are several other groups in the experiment and other measures with which we are not concerned here. One measure of outcome was a score or "grade" given by the instructor on a standard flight test in the airplane. To check upon this score, each subject was given a flight test also by a CAA flight examiner. The problem is to determine how closely the ten pairs of scores agree.

One method of doing this is to compute the *covariance*. This we shall first do in a direct manner in order to see clearly the concepts involved, though in practical work routine computations may be made similar to the "short-cut" methods of Chapters 3 and 4.

The covariance of the variables X_1 and X_2 in the *population* may be defined as

$$\text{cov} = (\Sigma x_1 x_2)/N \tag{8.1}$$

but, since we are dealing with a sample, we should use an *estimate* of the population covariance. An *unbiased estimate* of the covariance is

$$\text{cov} = (\Sigma x_1 x_2)/(n - 1) \tag{8.2}$$

The denominator, $n - 1$, represents "*degrees of freedom* and is required for the same reason that $n - 1$ is required in formula 7.3 for s^2, the unbiased estimate of the variance. Note that, if the sample is large, the difference between n and $n - 1$ is negligible.

In columns 3 and 4 of Table 8.1 we have computed the deviation scores of the two variables for each of the ten subjects. These are found simply by subtracting the respective means, $\bar{X}_1 = 67.4$ and $\bar{X}_2 = 66.0$

from the raw scores of columns 1 and 2. On a calculator we compute the product of x_1 and x_2 for each pair cumulatively to get the sum of *product deviations*. Dividing by $(n - 1) = 9$, we have the cov $= 77.0/9 = 8.56$.

TABLE 8.1

FLIGHT TEST GRADES BY TWO DIFFERENT EXAMINERS ON TEN STUDENTS IN A PILOT TRAINING EXPERIMENT*

Student	Raw Scores		Deviation Scores		Standard Scores	
	X_1	X_2	x_1	x_2	z_1	z_2
	(1)	(2)	(3)	(4)	(5)	(6)
a	70	69	+2.6	+3.0	+0.59	+0.86
b	72	65	+4.6	−1.0	+1.04	−0.29
c	66	67	−1.4	+1.0	−0.32	+0.29
d	67	69	−0.4	+3.0	−0.09	+0.86
e	67	63	−0.4	−3.0	−0.09	−0.86
f	74	68	+6.6	+2.0	+1.49	+0.57
g	71	71	+3.6	+5.0	+0.81	+1.43
h	62	66	−5.4	0.0	−1.22	0.00
i	60	62	−7.4	−4.0	−1.67	−1.14
j	65	60	−2.4	−6.0	−0.54	−1.71
Sums	674	660	0.0	0.0	0.00	0.01
Sums of squares	45,604	43,670	176.4	110.0	8.99	8.98

* X_1 = grades by CAA examiner; X_2 = grades by instructor.

The covariance as defined by equation 8.1 may be recognized as a "mean," the *mean-product-deviation*. As such, any covariance, or estimate of it, is comparable to any other as far as differences due to size of sample or size of population are concerned. In the above form, however, it lacks general meaning because of the limitless choice of units of measurement. The agreement between annual income and expenditures for education in ten states may be no more than that for the ten scores of Table 8.1, yet the covariance estimate would certainly be a figure many times larger than our covariance.

We are already familiar with a method of handling this type of situation. As in Section 4.7, we may express all measures in *standard score* form. This requires computing the two standard deviations. Using equation 7.3 and the data of Table 8.1, we obtain

$$s_1^2 = 176.4/9 = 19.60; \quad s_1 = 4.43$$
$$s_2^2 = 110.0/9 = 12.22; \quad s_2 = 3.50$$

Defining $z = x/s$, we divide the deviations of columns 3 and 4 by s_1 and s_2, respectively, to get z_1 and z_2 in columns 5 and 6. On the calculator we find $\Sigma z_1 z_2 = 4.95$. This divided by $n - 1$ should be free of the magnitude of the scores. The result will be a "pure number." When we follow this procedure with different kinds of paired measures, the results should be standard.

This mean-product-deviation of standard scores is the *correlation coefficient*, sometimes called the "Pearson-product-moment-correlation coefficient," after Karl Pearson. The population value is designated ρ, following the familiar convention of using Greek letters for population parameters. From a sample, the notation to be used is r. Therefore, we may write the correlation between variable 1 and variable 2 as

$$r_{12} = (\Sigma z_1 z_2)/(n - 1) \tag{8.3}$$

In our example the correlation between X_1 and X_2 is $r_{12} = 4.95/9 = .55$. We might have arrived at this result more quickly by "correcting for s_1 and s_2" at one stroke. The covariance divided by the product of s_1 and s_2 is the same as equation 8.3. From equation 8.3

$$\begin{aligned} r_{12} &= \left(\sum \frac{x_1 x_2}{s_1 s_2}\right) \Big/ (n - 1) \\ &= \frac{\Sigma x_1 x_2}{(n - 1) s_1 s_2} \\ &= \frac{\mathrm{cov}_{12}}{s_1 s_2} \end{aligned} \tag{8.4}$$

Using equation 8.4, we verify our earlier computation to find that $r_{12} = 8.56/(4.43)(3.50) = .55$. In passing, we observe that

$$\mathrm{cov}_{12} = s_1 s_2 r_{12} \tag{8.5}$$

which is another way of defining the covariance. Of course, in the population, with which we rarely deal, the covariance would be $\sigma_1 \sigma_2 \rho_{12}$.

8.2 THE PRODUCT MOMENT CORRELATION

In the previous section we have developed the correlation coefficient (equations 8.3 and 8.4) from the concept of covariance. We may use other points of reference as a means of comprehending correlation. As a matter of fact, we should be able to perceive of correlation from several different standpoints to understand its meaning thoroughly so that we may use it and interpret it intelligently.

The limits of the correlation coefficient are -1 and $+1$. In the example of Table 8.1, if the original scores in columns 1 and 2 had been *identical*, the z-scores in columns 5 and 6 would have been identical. The formula for r (equation 8.3) would then be equivalent to the squaring of the z's for either the first variable or the second one, summing and dividing by $(n-1)$. This you will recognize simply as the variance of one of the z scores. In other words, equation 8.3 would become

$$\frac{\Sigma z^2}{n-1} = \frac{\Sigma x^2}{s^2(n-1)} = \frac{s^2}{s^2} = 1$$

If the two variables of columns 1 and 2 in Table 8.1 were identical but in reverse order, that is, the highest in the first column had the lowest score in the second column and so on, in columns 5 and 6 the standard scores would be identical in numerical value but opposite in sign. In this case the result of applying formula 8.3 is numerically the same as the previous illustration, but each product is negative; therefore, their sums are negative, and the correlation is -1.

On the other hand, if paired with any value of a score in the first variable, most any of the values occur in the second, the relationships between the z scores in 5 and 6 may be expected to be quite haphazard. About as many of them would be of unlike signs as of like signs and the sums of the products of those with deviations in the same direction might, therefore, be expected to be equal to those with deviations in opposite directions. The numerator of equation 8.3 as a consequence of summing the product would be at or near zero. This represents an *absence* of relationship between the two variables.

A common formula used in computing correlation is equation 8.4. A slight modification yields

$$r_{12} = \frac{\Sigma x_1 x_2}{\sqrt{(\Sigma x_1^2)(\Sigma x_2^2)}} \tag{8.6}$$

In Chapter 4 we developed formula 4.14 for computing the sums of squares of deviations, using gross scores. This was $\Sigma x^2 = \Sigma X^2 - (\Sigma X)^2/n$. It is easy to derive the following similar formula for computing sums of products:

$$\Sigma x_1 x_2 = \Sigma X_1 X_2 - (\Sigma X_1)(\Sigma X_2)/n \tag{8.7}$$

We may thus write equation 8.6 as follows:

$$r_{12} = \frac{\Sigma X_1 X_2 - (\Sigma X_1)(\Sigma X_2)/n}{\sqrt{[\Sigma X_1^2 - (\Sigma X_1)^2/n][\Sigma X_2^2 - (\Sigma X_2)^2/n]}} \tag{8.8}$$

or

$$r_{12} = \frac{n\Sigma X_1 X_2 - (\Sigma X_1)(\Sigma X_2)}{\sqrt{[n\Sigma X_1^2 - (\Sigma X_1)^2][n\Sigma X_2^2 - (\Sigma X_2)^2]}}$$

There are many ways in which modifications may be made of this by algebraic manipulation.

From the data of columns 1 and 2 we may substitute in equation 8.8 as follows:

$$r_{12} = \frac{44{,}561.0 - 44{,}484.0}{\sqrt{(45{,}604.0 - 45{,}427.6)(43{,}670.0 - 43{,}560.0)}}$$

$$= \frac{77.0}{\sqrt{(176.4)(110.0)}} = \frac{77.0}{139.3} = .55$$

producing the same result as previously.

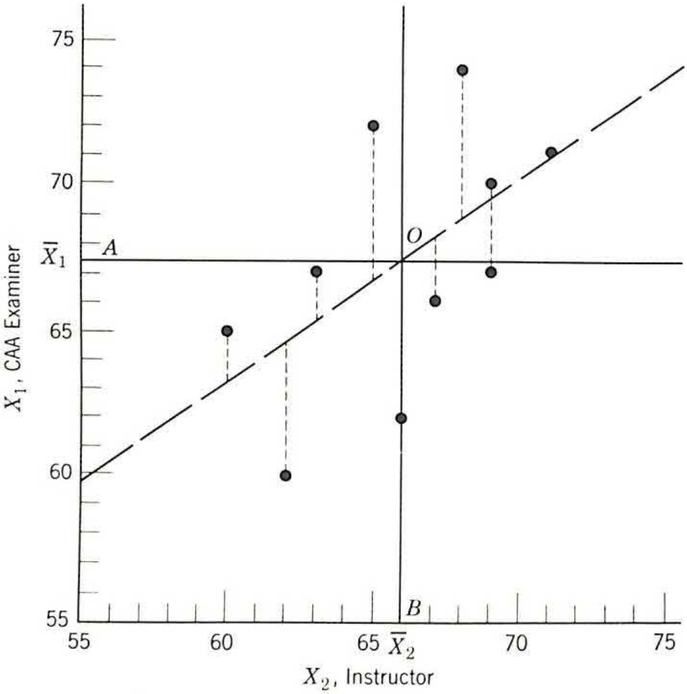
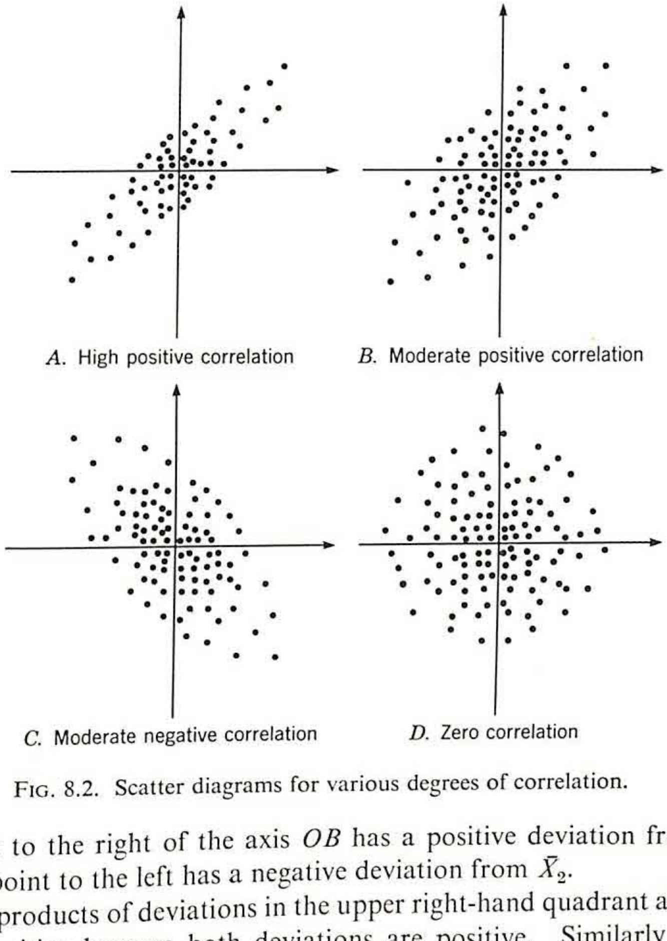


FIG. 8.1. Scatter diagram of flight test grades.

For larger samples than the one we have been discussing, it is usually advantageous to arrange data grouped in a correlation table, as is explained in Section 8.7. In any event, it is often desirable to arrange data either by table or graphically so that the *bivariate distribution* may be examined. For small distributions, such as our example, a *scatter diagram* is constructed as illustrated in Fig. 8.1. One of the variables is measured along the vertical axis, and another one on the horizontal axis. Each pair of measures form the coordinates for a point in the diagram.

In drawing a scatter diagram a convenient set of axes is chosen for plotting the points. A new set may then be drawn at distances of the *means* from the respective original axes. Any point on the line *AO* is thus a distance of 67.4 from the origin in X_1 units. Any point on the line *BO* is 66.0 units from the original vertical axis. Thus any point representing a score which is above *AO* has a positive deviation from the mean, and any point below *AO* has a negative deviation from the mean, $\bar{X}_1$. Similarly, any point to the right of the axis *OB* has a positive deviation from $\bar{X}_2$, and any point to the left has a negative deviation from $\bar{X}_2$.


A. High positive correlation

B. Moderate positive correlation

C. Moderate negative correlation

D. Zero correlation

FIG. 8.2. Scatter diagrams for various degrees of correlation.

All the products of deviations in the upper right-hand quadrant about *O* will be positive because both deviations are positive. Similarly points located in the lower left-hand quadrant will have positive products because both deviations are negative. Points located in the other two quadrants will have negative product deviations because, in either case, one of the deviations will be negative and the other positive. It may be seen from Fig. 8.1 that the points in the upper right-hand and lower left-hand quadrants are farther from the origin, *O*, as a whole, and there are more of them. It is to be expected, therefore, that the sum of products in these two quadrants will be greater than in the other two quadrants, explaining the fact that the sum of the products of deviations which we found in Section 8.1 to be 77.0 is positive. The positive sum of products of deviation is reflected in the positive correlation coefficient of .55.

The shape and "scatter" of such a diagram reveals a considerable amount of information. In Fig. 8.2 are illustrations of four scatter diagrams. *A* represents a high positive correlation characterized by the location of points along a thin elliptical area. A moderate positive correlation is shown in chart *B*; it is characterized by an elliptical shape with most points located in the upper right-hand and lower left-hand quadrants. Chart *C* is indicative of a similar degree of correlation though negative. There is a predominance of points in chart *C* in the upper left-hand and lower right-hand quadrants, those in which product deviations are negative. A diagram whose scatter about the origin is the same in all directions reveals *zero* correlation, illustrated by chart *D* in Fig. 8.2, the pattern of which is very similar to the "random" pattern of shot from a "bull's eye" with a shotgun.

One method of studying the differences between the four charts of Fig. 8.2, or for that matter, differences between any two correlation charts is to study a line which best fits the swarm of points in each diagram, and the extent to which the swarm of points fits the line. For this purpose we now consider the *regression line*.

8.3 REGRESSION AND PREDICTION

Let us now see how we "fit" a line to the points in Fig. 8.1 and find an equation for this line. There are several possible methods, but the one which most neatly fits correlation theory, and the one which is most commonly used in statistical work, assumes (*a*) that the best fitting line is a *straight line* and (*b*) that it is "best" when the sum of the squared deviations from the line is a minimum.

Let us assume that the broken straight line in Fig. 8.1 satisfies this latter condition, known as *least squares*. Then the sum of squares of deviations, that is, vertical distances of plotted points from this line, is a minimum. Figure 8.3 will assist us in recognizing the different values to consider. It shows the major axes for raw scores and in addition the two axes at $\bar{X}_1$ and $\bar{X}_2$, respectively. Also it shows a broken line representing the regression line. Point *P* in the chart represents one pair of values (X_2, X_1) in a scatter diagram.

We will consider X_1 the dependent variable and X_2 the independent variable, and seek a line which will give a good "prediction" of an individual X_1 score on the basis of the X_2 value only. Hence we are interested for the moment, in the vertical or X_1 value of this individual, and the chart shows the different line segments or values which we should consider.

First, we may break the total score for this individual X_1 into the two components $\bar{X}_1$ and x_1, the mean and the deviation of the score from the mean in the first variable, respectively. The broken line intersects the ordinate of point P at a distance from the base line which we shall call $\hat{X}_1$. This is the *regression* value or "predicted value" of the first variable for the individual whose scores are represented in the diagram. The remainder of the ordinate, X_1, is $X_1 - \hat{X}_1 = x_{1.2}$. As seen on the chart, it is the deviation of the actual raw score from the regression value. It represents the amount by which our line *misses* the true value of X_1. Also $\hat{X}_1$ deviates $\hat{x}_1$ from the mean, $\bar{X}_1$.

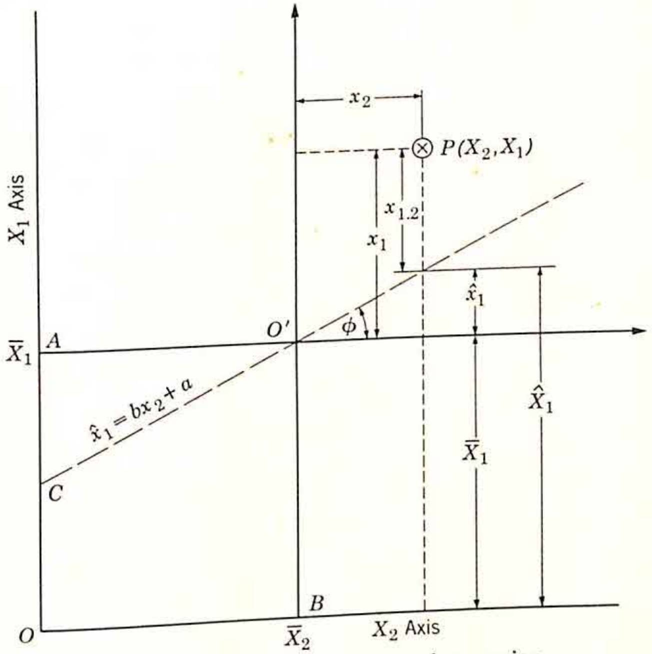


FIG. 8.3. Relation of scores and regression.

To summarize, there are three *gross score* values in the dependent variable:

(*a*) The actual score of an individual in the dependent variable, X_1.

(*b*) The mean, $\bar{X}_1$, of all the X_1 values.

(*c*) The ordinate of a point on the regression line which corresponds to an individual X_2 score—the regression value or predicted value, $\hat{X}_1$.

Then there are three *deviations*:

(*a*) The deviation of X_1 from the mean, $\bar{X}_1$.

(*b*) The deviation of the regression value from the mean.

(*c*) The deviation of X_1 from the regression value.

The third deviation, $x_{1.2}$, will be of special interest to us. We will call it the *residual* (that part of the X_1 score "left over," that is, not accounted for by the regression value, $\hat{X}_1$).

Of course, some of the points in a scatter diagram will be below the line representing the mean, or below the regression line, and these deviations will be negative. It can be shown that if the sum of squares of all the *residuals* is to be a minimum, they will sum to zero. That is, the positive deviations and the negative deviations from the regression line must be equal. We will recognize that the residuals, as deviations around the line, are being treated just as we treated deviations from a mean in computing variances.

From the diagram it may be seen that $x_1 = \hat{x} + x_{1.2}$. Summing for all individuals in the sample, we have

$$\Sigma x_1 = \Sigma \hat{x}_1 + \Sigma x_{1.2}$$

Now if Σx and $\Sigma x_{1.2}$ are zero, $\Sigma \hat{x}$ also equals zero. For that reason we know that $\bar{X}_1$ is the mean of the $\hat{X}_1$ values, or regression values.

A straight line is of the form

$$\hat{X}_1 = bX_2 + a \tag{8.9}$$

where *b* is the "slope" of the line and *a* is the intercept on the vertical axis. From geometry and trigonometry we know that the slope *b* is equal to the tangent of the angle ϕ between the line and the horizontal axis. It is easily proved that the regression line passes through the intersection of the two lines representing the means, that is, lines *AO* and *BO*, respectively. We may thus deal with residuals from the new axes with origin at *O*, the axes representing means of the two variables.

Since *b* is the slope of the line, we may now write

$$\hat{x}_1 = bx_2 \tag{8.10}$$

This is the *regression equation in deviation form*. It is desired to find *b* such that the sum of squares of the residuals will be a minimum. Any single residual, as seen from Fig. 8.3, may be written as follows:

$$x_{1.2} = x_1 - \hat{x}_1 = x_1 - bx_2$$

Thus the sum of squares of these residuals is

$$\Sigma(x_1 - bx_2)^2 = \Sigma x_1^2 - 2b\Sigma x_1 x_2 + b^2 \Sigma x_2^2$$

For those who have studied calculus it is simple to find the minimum value of this expression by differentiating with respect to b and setting the derivative equal to zero. The derivative is

$$-2\Sigma x_1x_2 + 2b\Sigma x_2^2$$

Setting this equal to zero and solving for b, we find that the value of b, for which the sum of squares of residuals is a minimum is

$$b = (\Sigma x_1x_2)/\Sigma x_2^2 \tag{8.11}$$

In order to get this into raw score terms, we may now write equation 8.10 in the following form:

$$(\hat{X}_1 - \bar{X}_1) = b(X_2 - \bar{X}_2)$$

Rearranging, this may be then written as follows:

$$\begin{aligned}\hat{X}_1 &= b(X_2 - \bar{X}_2) + \bar{X}_1\\ &= bX_2 + (\bar{X}_1 - b\bar{X}_2)\end{aligned}$$

Comparing this with equation 8.9 we see that

$$a = \bar{X}_1 - b\bar{X}_2 \tag{8.12}$$

These values of a and b give the line which makes the sum of squares of the residuals a minimum.

We might have considered a line for predicting X_2 from X_1. If we used the same diagram, the dependent variable would be measured horizontally, but the theory would be the same. As a matter of fact, we could make either variable the *dependent* variable, and choose whichever axis we wish for either, and this would make no difference in our analysis.

There are thus two regression lines. They coincide only when correlation is perfect. To distinguish between the slopes of these lines we will use b_{12} to represent the *regression coefficient* for the regression of X_1 on X_2 (the regression for predicting X_1 from X_2 values). In like manner, b_{21} will be the regression coefficient for the equation predicting X_2 values from X_1. In summary these equations are as follows:

In deviation form,

$$\begin{aligned}\hat{x}_1 &= b_{12}x_2\\ \hat{x}_2 &= b_{21}x_1\end{aligned} \tag{8.13}$$

In raw score form,

$$\begin{aligned}\hat{X}_1 &= b_{12}X_2 + a_{12}\\ \hat{X}_2 &= b_{21}X_1 + a_{21}\end{aligned} \tag{8.14}$$

The regression coefficients are

$$\begin{aligned} b_{12} &= (\Sigma x_1 x_2)/\Sigma x_2^2 \\ &= r_{12}\frac{s_1}{s_2} \\ b_{21} &= (\Sigma x_1 x_2)/\Sigma x_1 \\ &= r_{12}\frac{s_2}{s_1} \end{aligned} \tag{8.15}$$

Note that

$$\begin{aligned} b_{12} &= (\Sigma x_1 x_2)/\Sigma x_2^2 \\ &= \frac{\text{cov}}{s_2^2} = \frac{\text{cov}\, s_1}{s_1 s_2 s_2} \end{aligned}$$

and, from equation 8.4,

$$b_{12} = r_{12}\frac{s_1}{s_2}$$

Similarly

$$b_{21} = r_{21}\frac{s_2}{s_1}$$

In order to find the regression coefficients of the regression lines for the data of Table 8.1 we proceed as follows:

$$b_{12} = 77.0/110.0 = .700, \text{ and } b_{21} = 77.0/176.4 = .437$$

Since we have already computed the correlation coefficient, an alternative computation is:

$$b_{12} = .55\,\frac{4.43}{3.50} = .700, \text{ and } b_{21} = .55\,\frac{3.50}{4.43} = .435$$

We see that the results agree except for rounding errors.

The intercepts are:

$$\begin{aligned} a_{12} &= \bar{X}_1 - b_{12}\bar{X}_2 \\ a_{21} &= \bar{X}_2 - b_{21}\bar{X}_1 \end{aligned} \tag{8.16}$$

For the data of Table 8.1 we find that

$$a_{12} = 67.4 - (.70)(66.0) = 21.2$$

and

$$a_{21} = 66.0 - (.44)(67.4) = 36.3$$

The regression equations are hence

$$\hat{X}_1 = .70X_2 + 21.2$$

and

$$\hat{X}_2 = .44X_1 + 36.3$$

It is of interest to note that the correlation coefficient is the square root of the product of the two regression coefficients. From equation 8.15 it is readily seen that

$$r_{12} = \sqrt{b_{12}b_{21}}$$

The values of *b* and *a* are "statistics," that is, they are sample values and subject to sampling variation. We shall discuss in the next chapter the sampling distributions of regression coefficients and of correlation coefficients. Following our usual system of notation, we could write the regression equation based on *the entire population* as follows:[1]

$$\hat{X}_1 = \beta_{12}X_2 + \alpha_{12} \tag{8.17}$$

The connotation of the term *regression* is usually not really appropriate. This term was introduced by Galton when he observed that the deviation of mean height of sons was less than the deviation of mean height of fathers, a tendency for biological characteristics to "regress toward the mean." It is much more meaningfúl to think of the regression line as representing the "line of best fit" under the condition of least squares which we have described. In addition we should think of it as a line which permits us to make "the best prediction" or the best estimate of values of one variable from another.

A third meaning we should get from regression relates directly to the regression coefficient. As we have seen, it is the slope of the line. As such it represents the *rate of change* in the dependent variable per unit of change in the independent variable. It is further of significance to note that if variables are measured in standard deviation units, each regression coefficient is identical to the correlation coefficient (equation 8.6). In other words, the regression equations for the standard scores are as follows:

$$\hat{z}_1 = r_{12}z_2, \text{ and } \hat{z}_2 = r_{12}z_1 \tag{8.18}$$

In this case the correlation coefficient is the slope of the regression line. When the correlation is +1.0 the slope is 1.0, and the angle whose tangent is 1.0 is an angle of 45°.

Still another method of understanding correlation and regression is in terms of *how well* the corresponding regression line fits the swarm of points. We shall, therefore, examine the amount by which the line misses the points, by a study of residuals.

[1] However, we will not use this notation because of the fact that α and β have been used to represent other statistical measures.

8.4 THE RESIDUAL VARIANCE

Let us now turn to the data of Table 8.1. We repeat the raw scores in Table 8.2, and in addition we show the predicted values, $\hat{X}_1$, based on the regression of X_1 on X_2.[1] We also show the residual, the difference between the actual X_1 value and the predicted value, in the fourth column of Table 8.2. Each figure in column 3 was computed by means of the regression equation. For instance, for student a, whose instructor grade was 69, we find $\hat{X}_1 = (.70)69 + 21.2 = 69.5$.

TABLE 8.2

FLIGHT TEST GRADES, PREDICTED VALUES OF X_1 AND RESIDUALS

Student	Raw Scores		Estimated Value of X_1	Residual
	X_1	X_2	$\hat{X}_1$	$x_{1\cdot2}$
	(1)	(2)	(3)	(4)
a	70	69	69.5	+0.5
b	72	65	66.7	+5.3
c	66	67	68.1	−2.1
d	67	69	69.5	−2.5
e	67	63	65.3	+1.7
f	74	68	68.8	+5.2
g	71	71	70.9	+0.1
h	62	66	67.4	−5.4
i	60	62	64.6	−4.6
j	65	60	63.2	+1.8
Sums	674	660	674.0	0.0
Sums of squares	45,604	43,670	45,481.5	122.5

After completing and checking the computations of each of the ten regression values we add them and find that the sum is 674. This sum divided by 10 is 67.4. This result verifies our earlier theoretical observations that the mean of the regression values is the same as the mean of the actual scores.

[1] As indicated in Section 8.3, there may be interest in two sets of *predicted* values, and two sets of residuals. The present discussion treats only the one as the principles apply equally to both.

The figures in column 4, the residuals, are each obtained by subtracting the column 3 figure from the corresponding column 1 figure. As we observed in our theoretical discussion, these sum to zero. At the foot of column 4 we have the sum of the squares of these residuals, 122.5. According to Section 8.3 this is the *least* sum of squares of deviations from a line which could be obtained, for the regression line is the "least squares" line. Now if we divide this by the proper number of degrees of freedom we will have a mean-squared-deviation, a *variance.*

We are dealing with sample data; therefore, we do *not* divide by 10, as we did *not* divide by 10 when computing our other variances and our covariance. The appropriate divisor in each case was $n - 1$. In the present instance, also, working with the residual variance, we must ascertain the appropriate number of "degrees of freedom" in order that the result will be an unbiased estimate of the population variance. We shall, therefore, digress for a moment to discuss degrees of freedom.

In our discussion of the t distribution in Chapter 7, we used $n - 1$ as the denominator in estimates of variance and hence in entering the t table. Of course, the sample of n items has n degrees of freedom. If the *population* mean were known, n deviations from it could be computed and the sum of squares would furnish an unbiased estimate of the population variance when divided by n. When the population mean is unknown, the n degrees of freedom may usefully be thought of as consisting of one degree of freedom which measures the deviation of the sample mean from the population mean (and is unknown), and $n - 1$ degrees of freedom among the members of the sample. It is easy to see that there are only $n - 1$ degrees of freedom in the n deviations from the *sample* mean. This follows from the fact that, though $n - 1$ of the deviations might have any values whatever, as soon as their values are fixed so is the value of the nth so that their total will be zero, as it must. We say that we lose a *degree of freedom* by estimating the mean, or that the mean represents a restriction on the degrees of freedom of the variables under discussion.

Again considering deviations from the regression line, we see that two statistics have been computed from the n observations. They are a_{12} and b_{12} of the regression formula. Therefore, the n deviations from the regression line have $n - 2$ degrees of freedom.

In a universe we might designate the residual variance, sometimes called the *variance of estimate*, as $\sigma^2_{1.2}$, but the sample estimate will be designated as $s^2_{1.2}$. We may now write

$$s^2_{1.2} = \frac{\Sigma(X_1 - \hat{X}_1)^2}{n - 2} = \frac{\Sigma x^2_{1.2}}{n - 2} \tag{8.19}$$

In our example we find $s^2_{1.2} = 122.5/8 = 15.31$. Extracting the square root we find that $s_{1.2} = 3.91$. This is the estimate of the standard deviation of the population of *residuals*, the deviations from the population regression line. It is called the *standard error of estimate.*

It is of importance to recall that in Section 8.1 we found the total variance of raw scores in the first variable to be $176.4/9 = 19.60$. By comparison with the variance of residuals, 15.31, we can tell how much regression *reduces* variability in X_1. We may consider the operations leading to column 4 of Table 8.2 to be removing part of each X_1 by means of the regression line. The regression values of column 3 have a variance which is *that variance in X_1 which may be predicted by X_2.* The amount that is left is represented by the standard error of estimate or, more specifically its square, the residual variance, which shows how much of the variation in X_1 is *not* explained by the variable X_2.

8.5 COMPONENTS OF VARIANCE

In order to proceed with our discussion of the sums of squares involved in the correlation and regression problem, we will examine the correlation between the estimated values of X_1 and the residuals. In deviation form, this involves the product sum, $\Sigma\hat{x}_1 x_{1.2}$. Since $\hat{x}_1 = b_{12}x_2$, and $x_{1.2} = x_1 - b_{12}x_2$, we may write for any single product the following:

$$\begin{aligned}\hat{x}_1 x_{1.2} &= b_{12}x_2(x_1 - b_{12}x_2) \\ &= b_{12}x_1x_2 - b_{12}^2x_2^2\end{aligned}$$

Summing over observations in the distribution, we may write

$$\Sigma\hat{x}_1 x_{1.2} = b_{12}\Sigma x_1x_2 - b_{12}^2\Sigma x_2^2$$

Substituting for b_{12} its equivalent form (8.15), and simplifying, we reduce this to

$$\Sigma\hat{x}_1 x_{1.2} = \frac{(\Sigma x_1x_2)^2}{\Sigma x_2^2} - \frac{(\Sigma x_1x_2)^2}{\Sigma x_2^2} = 0 \tag{8.20}$$

Hence, *the correlation between the regression value and the residual in the dependent variable is zero.* This is another way of seeing that the residual is that part of the dependent variable which is not related to the independent variable.

You may verify this by computing the deviations of $\hat{X}_1$ in column 3 of Table 8.2 from their mean of 67.4. For student a this deviation $\hat{x}_1$ will be $+2.1$. These ten deviations, $\hat{x}_1$, multiplied by the corresponding residuals of column 4 in Table 8.2, and summed, yield zero. Of course if the sum of product deviations in the numerator of the correlation coefficient is zero, the correlation must be zero, so there is no need to compute the denominator of the correlation formula. (See formula 8.6.)

By definition (see Fig. 8.3)

$$x_1 = \hat{x}_1 + x_{1.2}$$

and

$$x_1^2 = \hat{x}_1^2 + 2\hat{x}_1x_{1.2} + x_{1.2}^2$$

Summing like values for all individuals in the distribution, we obtain

$$\Sigma x_1^2 = \Sigma\hat{x}_1^2 + \Sigma x_{1.2}^2 + 2\Sigma\hat{x}_1x_{1.2}$$

As we have seen in equation 8.20, the product sum of the last term in the foregoing is zero so that

$$\Sigma x_1^2 = \Sigma\hat{x}_1^2 + \Sigma x_{1.2}^2 \qquad (8.21)$$

This may be verified in the example of Tables 8.1 and 8.2. The sum of squares of deviations of raw scores from the mean in the first variable is shown in column 3 of Table 8.1 to be 176.4. We can find deviations of the estimated values of column 3, of Table 8.2, square them and sum them, or we can use a gross score formula and compute them to find that $\Sigma\hat{x}_1^2 = 45{,}481.5 - 45{,}427.6 = 53.9$. From column 4 of Table 8.2 we see that the sum of squares of the residuals is 122.5. Thus, $176.4 = 53.9 + 122.5$, verifying the identity (equation 8.21). The sums of squares of deviations of predicted values from $\bar{X}_1$, $\hat{x}_1 = b_{12}x_2$, may be computed by means of the following:

$$\Sigma\hat{x}_1 = \Sigma b_{12}^2x_2^2 = b_{12}^2\Sigma x_2^2 = \left(\frac{\Sigma x_1x_2}{\Sigma x_2^2}\right)^2\Sigma x_2^2 = \frac{(\Sigma x_1x_2)^2}{\Sigma x_2^2} \qquad (8.22)$$

Substituting in equation 8.21 and solving for $\Sigma x_{1.2}^2$, we have the following:

$$\Sigma x_{1.2}^2 = \Sigma x_1^2 - \frac{(\Sigma x_1x_2)^2}{\Sigma x_2^2} \qquad (8.23)$$

In our development of equation 8.21 we have demonstrated that the sum of squares *total* (of deviations from the mean of X_1) is equal to: (*a*) the sum of squares of deviations of the regression values from the mean of X_1 *plus* (*b*) the sum of squares of residuals.

This additive feature of "sums of squares" of deviation suggests the possibility that the variances are additive, because the numerators of variances are sums of squares. This is the case when the denominators for the variances are the same. However, we have noted in Section 8.4 that the denominator in the *total* variance in X_1 or in X_2 is $n - 1$, whereas the denominator for the residual variance is $n - 2$. This is necessary when we are working with a sample, which is most often the case, and when we are interested in *estimating* the universe value. If we deal with the entire population of measures, N, we may divide equation 8.21 by N. The result would be

$$\sigma_1^2 = \sigma_{\hat{x}_1}^2 + \sigma_{1.2}^2 \tag{8.24}$$

where $\sigma_{\hat{x}_1}^2 =$ the variance of predicted values. That is, in the universe, and virtually so in large samples, the total variance in the dependent variable consists of two parts: (1) the variance of the regression values, and (2) the residual variance. We may also write

$$1 = \sigma_{\hat{x}_1}^2/\sigma_1^2 + \sigma_{1.2}^2/\sigma_1^2$$

to show that the *proportion* of the total variance in X_1 due to regression plus the *proportion* due to residuals is equal to the total.

We turn our attention again to the sample situation, with particular reference to the relationship of these components of variance to the correlation coefficient. We will do this by means of a further examination of the *standard error of estimate.*

8.6 THE STANDARD ERROR OF ESTIMATE

The sum of squares of residuals, from equation 8.21, is equal to the total sum of squares minus the sum of squares of regression values,

$$\Sigma x_{1.2}^2 = \Sigma x_1^2 - \Sigma \hat{x}_1^2$$

Substituting from equation 8.22 we may express the residual sum of squares in terms of deviations in X_1 and X_2 as follows:

$$\Sigma x_{1.2}^2 = \Sigma x_1^2 - b_{12}^2 \, \Sigma x_2^2$$

Dividing by the appropriate degrees of freedom, we have

$$\frac{\Sigma x_{1.2}^2}{(n-2)} = \frac{(n-1)}{(n-2)} \left[\frac{\Sigma x_1^2}{(n-1)} - b_{12}^2 \, \frac{\Sigma x_2^2}{(n-1)} \right]$$

In terms of variances

$$s_{1.2}^2 = \frac{(n-1)}{(n-2)} [s_1^2 - b_{12}^2 s_2^2] \tag{8.25}$$

Referring to equation 8.15 and substituting for b_{12}^2, we have

$$\begin{aligned} s_{1.2}^2 &= \frac{(n-1)}{(n-2)} \left[s_1^2 - r_{12}^2 \, \frac{s_1^2 s_2^2}{s_2^2} \right] \\ &= \frac{(n-1)}{(n-2)} \, s_1^2 (1 - r_{12}^2) \end{aligned} \tag{8.26}$$

This is one formula for the *variance of estimate* from a sample. Its square root yields the *standard error of estimate*. The variance of estimate may also be computed from equation 8.23 and dividing by $(n - 2)$.

In Section 8.4 we directly computed the residual variance and standard error of estimate for the regression of X_1 on X_2 in Table 8.1. Formula 8.26 permits us to arrive at the same result without the necessity of computing each residual—squaring, summing, and dividing by degrees of freedom. Substituting the data from Table 8.1 in equation 8.26, the standard error of estimate agrees, except for rounding errors, as follows:

$$s_{1.2}^2 = (\tfrac{9}{8})19.60(1 - .302) = 15.39$$

$$s_{1.2} = 3.92$$

In statistical work with large samples the fraction in equation 8.26 is often disregarded, and the following *universe* relationship of parameters is used for the residual variance:

$$\sigma_{1.2}^2 = \sigma_1^2(1 - \rho_{12}^2), \tag{8.27}$$

or, for the standard error of estimate,

$$\sigma_{1.2} = \sigma_1 \sqrt{1 - \rho_{12}^2} \tag{8.28}$$

where ρ is the *population* correlation. From this we see that the *proportion* of total variance which is *residual* variance, in the dependent variable is equivalent to one minus the square of ρ,

$$\sigma_{1.2}^2/\sigma_1^2 = 1 - \rho_{12}^2$$

From equation 8.24 in the previous section we may show that

$$\sigma_{\hat{x}_1}^2/\sigma_1^2 = \rho_{12}^2$$

It follows that the correlation coefficient is directly interpretable in terms of proportions of variance in the dependent variable accounted for by the two variance components, the residual variance and the variance due to regression. The square of ρ is literally the proportion of variance in X_1 which is calculable by regression from X_2. The remainder, $1 - \rho^2$, of course, is the proportion of variance *not* explained by this regression.

Reference to equation 8.28 will show that the factor, $\sqrt{1 - \rho_{12}^2}$, is the ratio of the standard error of estimate to the standard deviation of the dependent variable. It is called the *coefficient of alienation*. It is a measure of *absence* of relationship between two variables. In fact, it is the correlation of the dependent variable X_1 with the residuals in X_1 (that part of the dependent variable *not* explained by regression).

A correlation of .50 corresponds to a coefficient of alienation of .87. This means that the standard error of estimate in predicting one variable from another is 87 percent of the standard deviation of the first variable if the correlation between them is .50. An aptitude test, for instance, which correlates .50 with achievement in first-year algebra, would *predict* achievement with a standard error of estimate only about 13 percent smaller than would result from using the mean as the prediction. It takes a correlation coefficient of at least .866 to reduce variability of prediction by as much as 50 per cent.

The foregoing procedure is frequently employed in the interpretation of correlations.

8.7 COMPUTING *r* FROM A CORRELATION TABLE

The *raw score* formula (8.8), or some modification of it, is most useful in computing r if the number of cases is small and a scatter diagram or *bivariate distribution* of the measures is not desired. Otherwise it is advantageous to use a "short-cut" method employing the principles of Sections 3.5 and 4.4. This method consists of measuring deviations (in both variables) in class interval units from an arbitrary reference point.

For this purpose special tabulation forms may be used, or a sheet of paper of convenient size ruled off in squares or cells to permit the tallying of frequencies in terms of both variables simultaneously.

The procedure is illustrated in Fig. 8.4. Note that this example uses only eight intervals in X_1 and six intervals in X_2. Ordinarily, to avoid errors due to such coarse grouping, a larger number of intervals should be used. (See Section 4.5.)

In our example there were three pairs of scores in Reading Comprehension and Reading Speed respectively, as follows: 28, 27; 30, 27; and 28, 29. Each of these individuals is in the 28–30 interval in the first variable (the vertical classification), and in the 27-29 interval in the second variable (the horizontal classification). This accounts for the three tallies in the cell for which scores of 28-30 in X_1 are also 27-29 in X_2. Tallies in the other cells account for the proper classification of the remainder of the 49 cases represented in the table.

After the tallying has been completed and checked, numbers are entered in cells for *frequencies* corresponding to the tally count.

We may now compute the correlation as follows:

$$r_{12} = \frac{\Sigma f x_1' x_2' - (\Sigma f x_1')(\Sigma f x_2')/n}{\sqrt{\Sigma f(x_1')^2 - (\Sigma f x_1')^2/n}\,\sqrt{\Sigma f(x_2')^2 - (\Sigma f x_2')^2/n}} \qquad (8.29)$$

It may be shown that equation 8.29 is equivalent to equation 8.6 since the numerator of 8.29 is $\Sigma x_1 x_2/u_1 u_2$ and the denominator is

X_2, Reading speed

X_1, Reading comprehension

	21-23	24-26	27-29	30-32	33-35	36-38	(1) f	(2) x'_1	(3) fx'_1	(4) $f(x'_1)^2$	(5) $\Sigma fx'_2$	(6) $\Sigma fx'_1 x'_2$		
31-33			I 1			I 1	2	4	8	32	3	12		
28-30			III 3	II 2	I 1		6	3	18	54	4	12		
25-27	I 1	II 2	IIII 4	III 3			10	2	20	40	−1	−2		
22-24	I 1	I 1	卌 5	IIII 4	I 1		12	1	12	12	3	3		
19-21		I 1	IIII 4	II 2			7	0	0	0	1	0		
16-18	I 1	II 2	III 3	I 1			7	−1	−7	7	−3	3		
13-15		II 2	I 1	I 1			4	−2	−8	16	−1	2		
10-12	I 1						1	−3	−3	9	−2	6		
(a) f	4	8	21	13	2	1	$n = 49$	☒	40	170	4	36		
(b) x'_2	−2	−1	0	1	2	3	☒							
(c) fx'_2	−8	−8	0	13	4	3	4							
(d) $f(x'_2)^2$	16	8	0	13	8	9	54							

FIG. 8.4. Computing r from a correlation table.

$\sqrt{(\Sigma x_1^2)(\Sigma x_2^2)/u_1u_2}$, where u_1 and u_2 are the class intervals. In equation 8.29 all product deviations and squares of deviations are in *interval units.* In passing we observe that this means that correlation is not changed by *linear transformations* of variables. That is, we may multiply or divide either or both variables by a constant and we may add a constant to or subtract a constant from either or both variables without changing r.

The data needed for formula 8.29 are obtained from rows *a–d* and columns 1–6 of the table in Fig. 8.4. The steps in the computation of a correlation problem by this method are as follows:

1. Sum the frequencies in each column, entering totals in row (*a*). This is the frequency distribution of X_2, Reading Speed, in the example of Fig. 8.4.

2. Similarly sum the frequencies of rows, entering totals in column 1, the frequency distribution of X_1. The grand totals of row *a* and column 1 each equal n.

3. Select arbitrarily a class interval in X_1 whose classmark will be the arbitrary reference point in X_1. With colored pencil or in some other manner accentuate the horizontal lines which set off the row corresponding to this interval. In Fig. 8.4 this is the row for the interval 19–21.

4. Similarly select arbitrarily an interval in X_2, accentuating the vertical lines which mark off the corresponding column. In Fig. 8.4 this is the column for the interval 27-29.

Note: Since any interval may be chosen, it is sometimes desirable to choose the lowest intervals in X_1 and X_2. This eliminates negative x' deviations and the extreme caution necessary otherwise in handling signed numbers.

5. Enter deviations from the reference point in class interval units for the first variable, x_1', in column 2, beginning with a zero entry in the row chosen as the reference interval. Be sure to mark negative deviations for intervals below the reference interval.

6. Similarly, beginning with zero for the chosen reference interval in X_2, enter deviations, x_2', in row *b*.

7. Multiply entries in columns 1 and 2 to obtain the fx_1' entries for column 3 and total. In our example, the total involves signs. We may check to see that $\Sigma fx_1' = 58 - 18 = 40$.

8. Similarly compute fx_2' in row *c* from products of entries in rows *a* and *b*. In Fig. 8.4 the algebraic sum of these entries is $\Sigma fx_2' = 4$.

9. Enter in column 4 the product of pairs of entries in columns 2 and 3 for each row. The total of these entries in column 4 is $\Sigma f(x_1')^2$. In our example this is 170.

10. Similarly derive entries in row *d* from entries in rows *b* and *c* for each column and sum to find $\Sigma f(x_2')^2$. This is 54 in the example.

11. Column 5 entries for each row are sums of deviations in the *second variable*, X_2. For instance, looking in the row for the interval 31–33 and row b, we see that there is one frequency whose x'_2 deviation is 0 and one whose deviation is 3. The sum of these deviations, 3, is entered in column 5 for the first row. In the second row, we see by reference to row b that there are three frequencies of 0, two of 1, and one of 2. The $\Sigma fx'_2$ entry for this row is thus 4. In the third row, the frequencies and deviations respectively, are: 1, -2; 2, -1; 4, 0; and 3, $+1$. The sum for this column is -1. Proceed in a similar manner through all rows to obtain an algebraic sum which should check with the total of row c.

12. In column 6 enter, for each row, the product of the column 2 and column 5 entries. Exercise extreme caution in keeping track of positive and negative signs in multiplying and in entering products. In our example, note that the entry for the third row is negative. Sometimes two columns are used for the $\Sigma fx'_1x'_2$ entries, one for the positive entries, and one for the negative ones in order to keep them separate. The total of column 6 is the algebraic sum of the product deviations in x'_1 and x'_2.

13. Substitute from the appropriate column and row totals the values required in formula 8.29. For the example of Fig. 8.4,

$$r = \frac{36 - (40)(4)/49}{\sqrt{170 - (40)^2/49}\,\sqrt{54 - (4)^2/49}}$$

$$= \frac{32.73}{\sqrt{137.35}\,\sqrt{53.67}} = .381$$

14. Compute the means of the two variables, using formula 3.7:

$$\bar{X}_1 = 22.45;\ \bar{X}_2 = 28.24.$$

15. Compute s_1 and s_2 from the two factors in the denominator of step 13. Note that the two radicands are the sums of squares of deviations from the respective means $\bar{X}_1$ and $\bar{X}_2$ *in class interval units*. It is necessary to convert these into *score units* by multiplying by the square of the class interval; see Section 4.4 and formula 4.13 modified to conform to formula 7.3. The result divided by $n - 1$ is the variance estimate. In our example u_1 and u_2 are each 3. Hence

$$s_1^2 = (9)(137.35)/48 = 25.75$$

Extracting the square root of this

$$s_1 = 5.07$$

The second variance

$$s_2^2 = (9)(53.67)/48 = 10.06$$

and

$$s_2 = 3.17$$

16. Determine the regression equations if these are required in the problem at hand. From equation 8.15 we may find the regression coefficients:

$$b_{12} = (.381)(5.08)/(3.17) = .61$$

and

$$b_{21} = (.381)(3.17)/(5.08) = .24$$

By means of equation 8.16 we may compute the intercepts:

$$a_{12} = 22.45 - (.61)(28.24) = 5.2$$
$$a_{21} = 28.24 - (.24)(22.45) = 22.9$$

The regression equations are:

$$\hat{X}_1 = .61X_2 + 5.2$$

and

$$\hat{X}_2 = .24X_1 + 22.9$$

17. Compute the standard error of estimate corresponding to each regression equation to be used. By means of equation 8.26 we find

$$s_{1.2}^2 = \frac{48}{47}(25.76)(1 - .145) = 22.5$$

and

$$s_{1.2} = 4.74$$

Similarly

$$s_{2.1}^2 = \frac{48}{47}(10.06)(1 - .145) = 8.78$$

and

$$s_{2.1} = 2.96$$

To find the residual variance and the standard error of estimate we may also use equation 8.23, substituting values in class interval units, multiplying by the square of the interval size, and dividing by $n - 2$. For example,

$$\begin{aligned} s_{1.2}^2 &= \Sigma x_{1.2}^2/(n-2) \\ &= (9)[137.35 - (32.73)^2/(53.67)]/47 \\ &= 22.47 \end{aligned}$$

Even though the data from the coarsely grouped scores of the example of this section do not justify it, two or more decimal places have been retained in the above computations to illustrate the procedure.

8.8 SOME CONSIDERATIONS IN INTERPRETING CORRELATION AND REGRESSION

There are a great many ways in which the correlation coefficient may be influenced. It is almost always wise to examine carefully the composition of measures to be correlated to see that there is not some common element which is accounting for some of the relationship. For instance, care should be exercised in the interpretation of correlations between ratios or indices. One common ratio in education is the I.Q., the ratio of mental age to chronological age. We must watch correlating this with some other type of ratio. The ratio of achievement age to chronological age, for instance, was once used as an educational quotient. In correlating these two quotients a high measure of relationship is inevitable because of the common denominator, chronological age. Similarly, we must guard against the correlation of one part of a test with the total score on the test (including the one part). Obviously the two measures, the part score and total score, will contain common elements which are perfectly correlated.

The interpretation of an observed correlation coefficient depends upon the nature of the bivariate distribution which it represents. Classical correlation theory is based on the *normal correlation* model which assumes (*a*) that each variable is normally distributed, (*b*) that X_1 is normally distributed for any specified X_2 (and vice versa), (*c*) that both regressions are straight lines, and (*d*) variance is homogeneous (that is, the criterion of homoscedasticity) in each variable throughout the range of the other. However, the definition of correlation in equation 8.6 makes no assumption of *normality* or *linearity of regression.*

The correlation and regression discussed in this chapter deals only with linear regression. One method of correlation analysis for nonlinear bivariate distributions is discussed in Chapter 13. It is enough to say here that, if a straight line is *not* the best fit, formulas such as 8.26 and 8.27 for residual variance based on linear regression will be *overestimates* in the sense that the residual variance would be smaller if a more suitable curve were fitted to the data. The correlation coefficient will thus *underestimate* the relationship of the two variables. In any event, it is important to take cognizance of the type of distribution with which we are working when assessing an observed correlation.

One of the most important considerations is the nature of the sample from which an r is observed. For instance, a sample of pupils from *all* elementary grades would most certainly yield a higher correlation between two educational variables than would a "homogeneous" population such as fourth graders only. The reason for this is that the wide range of scores on both tests in the first instance would tend to "string out" the pattern of the scatter diagram compared to the pattern for the homogeneous group. If the observations from which a correlation is computed were selected in some "non-random manner," no interpretation of r can can be made. In a tryout of a battery of proficiency tests for radar maintenance mechanics on a small group of subjects, very high intercorrelations, validity coefficients and reliability coefficients, were derived simply because the group was picked to represent "from the very poorest to the very best." The group was selected mostly from among the poorest and the best and few in between. In this instance the extreme cases produced a high covariance.

One common restriction on randomness of sampling in correlation occurs by selecting a group randomly from within a limited interval in one variable. Students frequently get into this situation in attempts to *match* groups in experiments. This problem is discussed in Chapter 11. The departure of this procedure from randomness in sampling should be clear from the discussion of Chapter 5.

As has been pointed out, our purpose in computing r is generally to infer something about a population ρ. For this reason, r almost always must be viewed as a sample statistic. We do, therefore, spend more time in the next chapter on the sampling distribution of r so that we may make statistical decisions of the type discussed in Chapter 7.

There are several advantages in *regression* analysis over *correlation* analysis. One is that it is usually sufficient for residuals in the dependent variable to be normally distributed and homoscedastic. Also, by means of regression functions other than the straight line, it is possible to avoid the assumption of linearity.

EXERCISES

1. Define:

Correlation coefficient.
Regression coefficient.
Covariance.
Mean-product-deviation.
Correlation table.
Scatter diagram.
Least squares.
Regression value.
Predicted value.
Residual variance.
Variance of estimate.
Standard error of estimate.
Coefficient of alienation.
Linear transformation.

2. Which of the following would you expect to be positively correlated, which negatively correlated, and which not related?

(*a*) Height and weight.

(*b*) Proficiency in a subject and the time it takes to solve a problem in the subject.

(*c*) Grade average in college and grade average in high school.

(*d*) Level of teacher's salaries of a school district and teacher turnover in the district.

(*e*) Intelligence and height.

(*f*) Cost per mile of operating a school bus and length of daily route covered.

(*g*) Size of high school and annual cost per pupil per year.

(*h*) The quality of the educational program of a school system and the drop-out rate of students.

(*i*) Cost of constructing a school building and total square feet of floor area.

(*j*) The age of a school bus and its trade-in value.

(*k*) Proportion of girls in the student body of a high school and proportion of graduates attending college.

(*l*) Mean annual rainfall in an area and the average age of the student body.

(*m*) An individual's age and his I.Q.

(*n*) An individual's age and the number of teeth he has.

(*o*) The number of unemployed and the volume of retail sales in a large city.

3. Make up a correlation work sheet similar to that in Fig. 8.4 on which to tabulate frequencies and to compute the correlation based on the 168 pairs of scores in Appendix A. Use the class intervals chosen for Ex. 1 of Chapter 3. When the tallies have been completed the frequency distribution as shown in Ex. 1 of Chapter 3 may be used to check the tabulation.

(*a*) Compute the correlation coefficient.

(*b*) Find the two regression equations.

(*c*) Find the standard errors of estimate.

4. From the results of the previous exercise:

(*a*) What is the best estimate of the writing score for a student who has a score of 75 on the California test?

(*b*) What is the best estimate of the writing score for a student whose mental maturity score is 55?

5. From the results of Ex. 3, answer the following questions:

(*a*) The *actual* writing score may be expected to vary from its regression estimate two-thirds of the time by as much as what amount?

(*b*) Approximately what proportion of the variance in writing score is not associated with variance in mental maturity?

(*c*) Describe a situation in which each of the two regression equations may be of practical interest.

6. Would you expect a correlation coefficient from the data of Appendix A, using gross scores and formula 8.8 to be the same as that computed in Ex. 3? Why?

7. The following are scores of 12 students on X_1, a first test, and X_2, a second test in a class in educational statistics:

Test	Student											
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
X_1	27	38	40	50	44	31	46	41	43	44	49	50
X_2	60	36	51	43	43	16	49	52	60	46	69	70

(*a*) What is the estimated score of a student who missed the second test, but who had a score of 39 on the first test?

(*b*) How good is this estimate?

8. The correlation between test A and test B for the large sample of students in the ninth grade is .70. The mean of test A is 125, and its standard deviation is 20. The mean of test B is 70, and its standard deviation is 10. One ninth-grade student is selected at random.

(*a*) If there is no score for him on either test, what is the best estimate you can make of what his score would bė on test A?

(*b*) What is the "standard error" of this estimate?

(*c*) Suppose that he has made a score of 65 on test B. What is the best estimate you can make of what his score would be on test A?

(*d*) What is the "standard error" of the estimate in *c*?

(*e*) Find the ratio of your answer for *d* to your answer for *b* and compare this with the coefficient of alienation. Explain.

9. The z scores of three pupils on test A are -1.6, 0.8, and 1.85. The correlation between test A and test B is .80. Estimate z scores for the three pupils on test B.

10. State-by-state figures on the annual per capita income and the average current expenditure per pupil correlate higher than .80 for the 48 states of the United States. Is this a sufficient basis for the argument that a state may increase the productive capacity of its people by increasing the level of educational opportunity? Explain.

11. A regression equation is found to be $\hat{X}_1 = 110X_2 + 5.0$ for predicting values of X_1 from observed values of X_2. To simplify the formula it has been decided to drop the constant term and reduce the regression coefficient to 100 so that values of the first measure will be estimated by $\hat{X}_1' = 100X_2$.

(*a*) Which will correlate highest with X_1, $\hat{X}_1$, or $\hat{X}_1'$?

(*b*) If the two prediction lines for the two prediction equations are plotted on a scatter diagram, which will best fit the points on the diagram?

(*c*) What is the relationship of the residuals from $\hat{X}_1$ to the residuals from $\hat{X}_1'$? To residuals from any other prediction of X_1?

12. What is the correlation coefficient from the two pairs of values: 23, 75; 84, 87, respectively? What is the correlation from the two pairs of values: 75, 23; 84, 87, respectively? Explain.

13. The correlation between mental age and height is found to be near zero for a large sample of boys of the same age. A correlation between mental age and height from a large sample of boys ranging in age from 6 to 16 was found to be significantly positive. How do you explain the difference in these results? How would you answer the question, "Are mental age and height correlated?"

14. From the data of Appendix G compute the correlation between the CTMM and the Physical Science Test scores for each of the following groups: (*a*) Male—college. (*b*) Male—noncollege. (*c*) Female—college. (*d*) Female—noncollege. What may account for differences among these four coefficients?

15. Under what circumstances would the product-moment correlation coefficient measure too *low* the relationship between two variables? Under what circumstances would it measure too *high* the relationship between the variables?

16. Estimate the coefficient of correlation between current expenditure per pupil in average daily attendance and value of taxable property per pupil in average daily attendance of school districts on the basis of the following information:

School District	Current Expense per Pupil	Property Valuation per Pupil
A	$188	$15,031
B	278	30,675
C	215	19,132
D	214	24,756
E	193	16,722
F	178	18,891
G	265	29,002

17. A research worker in a school system is interested in finding the correlation between two tests recorded on ninth-grade pupil personnel record cards. One test is an intelligence test, the other is a reading test. The intelligence test scores are percentiles from the test publisher's table of percentile norms for ninth-grade students. The reading test scores are linear transformations of raw scores based upon a system of norms similar to the Z score such that 50 is the norm for the ninth grade and the standard deviation is 10. Why is it not advisable to compute the correlation coefficient between the two test scores as recorded? What might be done to permit the computation of a more interpretable correlation?

18. Assume that the first ten pairs of scores in Appendix A are a random sample of pairs from a population. Compute the correlation between X_1 and X_2 from these ten pairs. Similarly compute the correlation between X_1 and X_2 from the next ten pairs of scores. How do you account for the difference between the two values of r_{12}?

19. A correlation of .45 was found between a midsemester test in an education class of 25 students in a large university and a number consisting of the last three digits of the students' identification cards. How do you account for this observed correlation?

20. The correlation of *mean item scores* of two forms of 50 items each of a test is .76. What is the correlation of the *total scores* of the two forms? (Total scores computed as sums of item scores.)

21. Prove that $\Sigma x_1x_2 = \Sigma X_1X_2 - (\Sigma X_1)(\Sigma X_2)/n$.

22. Prove that multiplying each observation of X_1 by a constant and each observation of X_2 by a constant or adding a constant to X_1 and adding a constant to X_2 will not change the value of r_{12}.

23. Prove that $r_{12}^2 = b_{12}b_{21}$.

REFERENCES

1. Edwards, Allen L., *Statistical Methods for the Behavioral Sciences*, New York, Rinehart and Co., 1954, Chapters 7 and 8.
2. Lindquist, Everet F., *A First Course in Statistics*, Revised Ed., Boston, Houghton Mifflin Co., 1942, Chapter 10.
3. Snedecor, George W., *Everyday Statistics Facts and Fallacies*, Dubuque, Iowa, Wm. C. Brown Co., 1950, Chapter 13.
4. Wilks, Samuel S., *Elementary Statistical Analysis*, Princeton, N. J., Princeton University Press, 1949, Chapter 13.

CHAPTER 9

Sampling Error in Correlation and Regression

It has been emphasized that correlation problems usually deal with samples from populations of pairs of measures. Hence a correlation coefficient is a *statistic*, and we may wish to test a hypothesis concerning the parameter, or population value. Similarly either of the two regression lines for a *sample* correlation chart or scatter diagram may be viewed as one of the large number of possible lines derived from similar samples. Although we may be concerned with only the one parameter, ρ, in analysis of the *correlation*, there are two parameters to consider in fixing the position of each of the two *regression lines*. Each regression equation (8.14) from a sample contains two *statistics*, one the *regression coefficient*, b, the other the *intercept*, a.

The general approach for testing statistical hypotheses and establishing confidence limits concerning ρ and the regression parameters is the same as that of Chapter 7, which dealt with sampling errors of the mean. There are, however, some special considerations to take into account when dealing with sampling error in correlation and regression analysis.

9.1 ERROR OF ESTIMATE VERSUS SAMPLING ERROR

There is one source of confusion which must be recognized at the very outset. We must carefully distinguish between *sampling error*, that is, the variation of a statistic such as r or b from sample to sample, and the *standard error of estimate*. The latter is purely a standard deviation of residuals from a regression line and is a measure of *error* of the line in predicting the dependent variable. It is not a standard deviation of the sampling distribution of either r or b.

9.2 THE BIVARIATE NORMAL DISTRIBUTION

Most of the theory concerning the sampling distribution of r is based upon a very special correlation model, the *normal correlation.* This model is a two-variable normal "bell-shaped" surface. Horizontal slices through it are ellipses, vertical slices through it are normal curves. For any value of one variable the distribution is normal for the other. A further feature is *homoscedasticity*, that is, for each variable the variance is the same throughout the range of the other variable. Still another characteristic is *linearity of regression.* In Section 8.3 our development of correlation assumed linearity of regression. Elaborate techniques have been developed for fitting curves to bivariate distributions and studying curvilinear regression. These are beyond the scope of this book.

There are methods of testing distributions in two variables to see if they meet the conditions of normality and linearity. The simplest of these, of course, is the visual examination of the correlation table or the scatter diagram.

In any event, the above conditions must be taken into account when an inference is to be drawn from an observed sample r. Moreover, the testing of statistical hypotheses about correlation is greatly facilitated under conditions of *normality*. There is very little exact knowledge about the sampling distribution of r except in the normal case. It is, therefore, very helpful to know whether or not the distribution with which we are working is one at least approximately meeting these requirements.

The sampling distribution of r is complicated because it depends upon the true correlation, ρ, of the universe, as well as the sample size, n. The sampling distribution of r is symmetrical for $\rho = 0$, but is extremely skewed for highly correlated (positive or negative) populations.

9.3 THE UNCORRELATED NORMAL POPULATION—LARGE n

If n is large, say, 50 or more, and $\rho = 0$, sample values of r are nearly normally distributed, with mean of zero and a standard deviation of

$$\sigma_r = 1/\sqrt{n-1} \tag{9.1}$$

Suppose that we wish to test the hypothesis that the sample of Section 8.7 is from an *uncorrelated normal population.* This is a null hypothesis. The hypothesis is that there is zero correlation in the population. The test of this hypothesis is sometimes considered a *test of independence*, but strictly speaking it is not, because of the assumption of linearity of regression. In other words, the two variables may have zero correlation, but be interdependent because of some other functional relationship. We compute $\sigma_r = 1/\sqrt{48} = .1443$. Using the normal curve, we proceed as in Section 7.5. Suppose that we choose $\alpha = .01$. Our test will be at the 1 percent level. In the present case $z = (r - \rho)/\sigma_r$ is assumed to be normally distributed. But since $H : \rho = 0$, substituting in equation 9.1, we compute

$$z = (r - \rho)/\sigma_r = .381/1443 = 2.64$$

A value of z of 2.58 is required at the 1 percent level. The observed z of 2.64 exceeds this. Therefore we reject the hypothesis of no correlation, at the 1 percent level.

9.4 AN EXACT TEST OF SIGNIFICANCE OF *r*

An exact method of testing the significance of an observed correlation makes use of the t distribution. As previously, the hypothesis is that the sample is from an uncorrelated bivariate normal population. It has been shown that in such a population

$$t = r\frac{\sqrt{n-2}}{\sqrt{1-r^2}}; \qquad \text{d.f.} = n - 2 \tag{9.2}$$

With this we find for the correlation of Section 8.7 that $t = (.381)(6.86)/(.925) = 2.83$. Reference to the table in Appendix E will show that for 47 d.f., $P(t \geq 2.83) < .01$. Hence, by this test also, r is significant at the 1 percent level.

The use of the t test, the method of equation 9.2, is clearly advised for testing the significance of r for samples of small n, particularly samples as small as that in Section 8.2. In that example, $r = .553$; $n = 10$; d.f. $= 8$; and $t = (.553)\ (2.83)/(.833) = 1.88$. For 8 d.f. the 5 percent level of t is 2.31. Therefore we *do not* reject the hypothesis of zero correlation in the population.

Suppose, on the other hand, that we observe a correlation of .80 from a sample of size 10. In this case we would reject the null hypothesis at the 1 percent level by the t test, but *not* by the method of Section 9.3. We would find $t = 3.77$ to exceed $t_{.995} = 3.36$, but $z = 2.40$ to be less than $z_{.995} = 2.58$.

Note that we have been making two-tailed tests. That is, our concern is whether an r, either *positive or negative*, by as much numerically as that observed, could have occurred by chance with probability of .05 or less, or .01 or less, whatever the level of significance is that has been chosen. If the conditions of the experiment are such that a one-sided test is to be used, the 5 percent and 1 percent levels for the t-test would be $t_{.95}$ and $t_{.99}$, respectively.

TABLE 9.1

ONE PERCENT AND 5 PERCENT LEVELS OF SIGNIFICANCE FOR THE CORRELATION COEFFICIENT*

Degrees of Freedom	Minimum Value of r To Be Significant	
	$p = .01$	$p = .05$
1	1.000	.997
2	.990	.950
3	.959	.878
4	.917	.811
5	.874	.754
6	.834	.707
7	.798	.666
8	.765	.632
9	.735	.602
10	.708	.576
12	.661	.532
14	.623	.497
16	.590	.468
18	.561	.444
20	.537	.423
22	.515	.404
24	.496	.388
26	.478	.374
28	.463	.361
30	.449	.349
40	.393	.304
50	.354	.273
60	.325	.250
70	.302	.232
80	.283	.217
90	.267	.205
100	.254	.195
150	.208	.159
200	.181	.138
500	.115	.088
1,000	.081	.062

* Without reference to sign of r. These are two-tailed significance levels.

By reference to tables of t, significance levels for various degrees of freedom may be determined by substituting the appropriate tabled values of t in equation 9.2 and solving for r. Table 9.1 was prepared in this manner. For 8 d.f. we see from this table that an observed r of at least

±.765 would be required at the 1 percent level and ±.632 at the 5 percent level. The r of .553 from Section 8.2 is thus seen as before to be too small to be significant (that is, to be reasonably sufficient grounds for rejection of $H : \rho = 0$).

The amount of error in using the method of Section 9.3 is suggested by the fact that it requires at the 1 percent level a correlation of .778 for 10 d.f. as compared with .708 in Table 9.1. For 20, 30, and 100 d.f., respectively, the required values of r are .563, .463, and .257, as compared with .537, .449, and .254 for the exact method of Table 9.1 or formula 9.2. At the 5 percent level, required values of r differ only in the third decimal for degrees of freedom greater than about 20.

9.5 FISHER'S z TRANSFORMATION

The above methods are useful only in testing the significance of r. They are not useful in testing a hypothesis other than zero correlation in the population, or in testing hypotheses about the differences between two correlations.

A transformation of r (that is, a function of r), z', has been shown by R. A. Fisher to have a sampling distribution which is approximately normal. This means that we can transform r into the corresponding z', and then use the normal distribution to make inferences about the population correlation ρ, of which r is an estimate, in the same way as the normal distribution was used in Chapter 7 in describing how well an observed $\bar{X}$ served to estimate μ.

The z transformation is as follows:

$$z' = \frac{1}{2}\log_e\left(\frac{1+r}{1-r}\right) \tag{9.3}$$

or

$$z' = 1.1513\log_{10}\left(\frac{1+r}{1-r}\right) \tag{9.4}$$

This may appear formidable to those not accustomed to using logarithms, but computations are not necessary if tables of z' equivalents of r are available. Such a table giving z' values for various values of r appears as Appendix F.

The standard deviation of z' is

$$\sigma_{z'} = \frac{1}{\sqrt{n-3}} \tag{9.5}$$

Suppose that we wish to test the hypothesis $H : \rho = .50$ for the population involved in the exercise of Section 8.7. If our test is to be made at the 5 percent level, we will reject, as we did in Chapter 7, if the observed normal deviate exceeds 1.96. We first convert the specified value of ρ and the observed r to z'. The mean of the approximately normal distribution of z' will be at the z' value corresponding to $\rho = .50$. Reference to the table in Appendix F will show this to be .549. The z' for the observed correlation of .381 is found by interpolation to be .401. Therefore, in terms of our hypothesis the observed r would represent a deviation from the mean in the z' distribution of .148. Substituting 49 for n in equation 9.5, we find the standard deviation of the distribution to be

$$\sigma_{z'} = 1/(6.78) = .147$$

In standard units the observed deviation would be $.148/.147 = 1.01$.[1] This is not sufficient grounds for rejecting the hypothesis.

A similar procedure, following that outlined in Chapter 7, is involved in using z' to establish *confidence intervals* for ρ. For instance, in a sample of forty school districts a correlation of .55 was found between average current expense per pupil and a rating scale on the adequacy of the educational program. The z' corresponding to the observed r is .618. From equation 9.5 we compute the standard deviation of z' for a sample of 40 and find it to be .164. Multiplying this by 1.96 will give us .321, half the proper interval *in z' units* for the 95 percent confidence limits. The limits *in z' units* are, therefore: $.618 - .321 = .297$, and $.618 + .321 = .939$. By means of Appendix F we convert these limits of z' into the r equivalents, .29 and .73. The interval includes all values of ρ which would be acceptable at the 5 percent level.

9.6 THE SAMPLING DISTRIBUTION OF THE REGRESSION COEFFICIENT

The sampling distribution of the regression coefficient is known when the residuals in the dependent variable are normally distributed and homoscedastic throughout the range of the independent variable. The standard error of the regression coefficient, estimated from a sample, is

$$\begin{aligned} s_{b_{12}} &= \frac{s_{1.2}}{\sqrt{\Sigma x_2^2}} \\ &= \frac{s_{1.2}}{s_2\sqrt{n-1}} \end{aligned} \qquad (9.6)$$

in terms of the sample standard error of estimate and the sample standard

[1] We are dealing here with the deviate of the normal curve which we have designated as z previously. It should not be confused with z', which is Fisher's transformation for r.

deviation of the independent variable. It may be computed directly from the "sum of squares" of deviations as follows:

$$s_{b_{12}} = \sqrt{\frac{\Sigma x_{1.2}^2}{(n-2)(\Sigma x_2^2)}} \tag{9.7}$$

From equation 8.26 this may be expressed as

$$s_{b_{12}} = \frac{s_1}{s_2}\frac{\sqrt{1-r^2}}{\sqrt{n-2}} \tag{9.8}$$

Following the reasoning of Chapter 7, we have the ratio

$$t = \frac{b-B}{s_b}; \qquad \text{d.f.} = n-2 \tag{9.9}$$

where B is the population parameter, distributed as Student's t.[1]

We may use this in the same manner as we used t in Chapter 7 to test hypotheses and to establish confidence limits for B. In the exercise of Table 8.1 we found $b_{12} = .70$; $s_{1.2} = 3.91$; and $\Sigma x_2^2 = 110$. Thus, from equation 9.6 we find

$$s_b = 3.91/\sqrt{110} = .373$$

For 95 percent confidence limits, with d.f. $= n - 2 = 8$, we find the critical value of $t = 2.31$ from Appendix E. The confidence limits for B are thus $.70 - (2.31)(.373) = -.16$; and $.70 + (2.31)(.373) = 1.56$.

To directly test the hypothesis that $B = 0$, equation 9.9 becomes

$$t = \frac{b-0}{s_b}; \qquad \text{d.f.} = n-2 \tag{9.10}$$

In the above exercise, this would be $t = .70/.372 = 1.88$, for which $P > .05$.

A re-examination of Section 9.4 will reveal that this is precisely the same t which was found in testing the hypothesis $H : \rho = 0$. This is no accident, for the t of equation 9.2 for testing the significance of r is equivalent to equation 9.10 for testing the significance of b. We may substitute equation 9.8 in 9.9 to prove this, remembering that $r = b_{12}s_2/s_1$. In any case it is important to note that knowing that r is significantly greater than 0 means that b_{12} and b_{21} are both significantly greater than 0.

[1] In using the symbol B for the parameter, we are departing from the usual rule of using Greek letters for the parameter. This is because we are reserving the Greek letter β for another special definition of regression coefficient which is more or less standard in the multiple regression methods of Chapter 13.

9.7 THE SAMPLING DISTRIBUTION OF THE REGRESSION CONSTANT

The regression equations computed from a sample contain two sample "statistics." In the previous section we have discussed methods of estimating the sampling variance of the regression coefficient. The other sample statistic is the intercept, $a_{12} = \bar{X}_1 - b\bar{X}_2$. The intercept is made up of the two components $\bar{X}_1$ and $b\bar{X}_2$, each of which is subject to variation from sample to sample.

It will be recalled that $s_{\bar{x}}^2 = s_x^2/n$ in a sample from a population for which there are measures on only the single variable X. It can be shown that the sampling variance of the mean of a dependent variable is the *residual* variance divided by the sample size, n. Therefore, the variance of the mean is

$$s_{\bar{x}_1}^2 = s_{1.2}^2/n \tag{9.11}$$

The second part of the sampling variance in the intercept, a_{12}, comes from $b\bar{X}_2$, which has a sampling variance because it involves the regression coefficient. In fact, the variance of this term is $\bar{X}_2^2 s_b^2$. The sum of these two variances is the variance of the intercept as follows:

$$s_{a_{12}}^2 = \frac{s_{1.2}^2}{n} + \frac{\bar{X}_2^2 s_{1.2}^2}{\Sigma x_2^2}$$

This may be expressed in a form more suitable for computation as

$$\begin{aligned} s_{a_{12}}^2 &= s_{1.2}^2\left(\frac{1}{n} + \frac{\bar{X}_2^2}{\Sigma x_2^2}\right) \\ &= s_{1.2}^2 \left[\frac{1}{n} + \frac{\bar{X}_2^2}{(n-1)s_2^2}\right] \end{aligned} \tag{9.12}$$

In the example of Section 8.3, we found $a_{12} = 21.2$. Given $\Sigma x_2^2 = 110$, $s_{1.2}^2 = 15.31$, $\bar{X}_2 = 66.0$, and $n = 10$, we substitute in equation 9.12 and find $s_a^2 = 607.81$ and $s_a = 24.7$. From this it is obvious that a_{12} is not significantly different from 0. We could use $s_a = 24.7$ to establish confidence intervals and to test other hypotheses concerning the population value, A_{12}, using

$$t = \frac{(a - A)}{s_a}; \qquad \text{d.f.} = n - 2 \tag{9.13}$$

9.8 SAMPLING ERROR OF A REGRESSION ESTIMATE

We have seen that a regression equation is subject to variation from sample to sample. This variation we may ascribe to sampling variation of a from the universe value A, and of b from the universe value B. Consequently, any estimate for a specified value, X'_2, of the independent variable, calculated from a sample regression equation should be expected to vary from an estimate based on the population regression equation which contains the parameters A_{12} and B_{12} instead of the sample statistics, a_{12} and b_{12}. The variance of $\hat{X}'_1$, estimated from a sample regression equation for a specified value, X'_2, is

$$s^2_{\hat{x}'_1} = s^2_{1.2}\left[\frac{1}{n} + \frac{(X'_2 - \bar{X}_2)^2}{\Sigma x_2^2}\right]; \qquad \text{d.f.} = n - 2 \tag{9.14}$$

It may be noted from equation 9.11 that the term involving n in equation 9.14 represents the contribution of sampling error in the mean, $\bar{X}_1$. The second term is the contribution of the sampling error of b. Furthermore, the second term is zero when the specified value, X'_2, of the independent variable is the mean $\bar{X}_2$. It increases as the square of the deviation of X'_2 from $\bar{X}_2$ increases. Consequently, the reliability of estimates from a regression equation may be represented by a "confidence band" as in Fig. 9.1. The farther the specified value of X_2 is from $\bar{X}_2$, the wider this band. For any specified value of the dependent variable, X'_2, confidence limits determining the band are

$$\hat{X}'_1 \pm t_0 s_{\hat{x}'_1}$$

where d.f. $= n - 2$, and t_0 is the appropriate t at the level of risk chosen.

The data for Fig. 9.1 were derived from the correlation problem of Section 8.7. We shall follow through the computations to see the application of equation 9.14.

9.9 CONFIDENCE BANDS FOR THE REGRESSION ESTIMATES

Suppose that our interest is in estimating Reading Comprehension, X_1, from Reading Speed, X_2, on the basis of the regression equation developed in Section 8.7. This equation, it will be recalled, was based on a sample of size $n = 49$. The equation was

$$\hat{X}_1 = .61X_2 + 5.2$$

A short-cut method of computing variances and covariances was used in Section 8.7, so we do not have the Σx_2^2 required in equation 9.14. However, we may derive it since $\Sigma x_2^2 = s_2^2(n - 1)$, and s_2^2 was found to be 10.06. Therefore, $\Sigma x^2 = (10.06)(48) = 482.9$. In section 8.7 we found $\bar{X}_2 = 28.24$, and $s_{1.2}^2 = 22.5$.

Substituting these values in equation 9.14, we have

$$s^2_{\hat{x}'_1} = 22.5\left[\frac{1}{49} + \frac{(X'_2 - 28.24)^2}{482.9}\right]$$

We may substitute in this various specified values for X'_2 and compute the sampling variance of the corresponding predictions, $\hat{X}'_1$.

For instance, the group in Fig. 8.4 whose Reading Speed was 24–26, are those for which we would assign the value $X_2 = 25$. In order to find the standard error of the prediction of X_1 from regression for this group, we substitute in the above equation to find

$$\begin{aligned} s^2_{\hat{x}'_1} &= 22.5\left[.02041 + \frac{(-3.24)^2}{482.9}\right] \\ &= 22.5(.02041 + .02174) \\ &= (22.5)(.04215) = .948 \end{aligned}$$

Extracting the square root, we find $s_{\hat{x}'_1} = .974$.

TABLE 9.2

DATA SHOWING CONFIDENCE LIMITS FOR REGRESSION ESTIMATES CHARTED IN FIG. 9.1

Specified Values, X'_2	Predicted Values $\hat{X}'_1$	Sampling Variance, $s^2_{\hat{x}'_1}$	Standard Error of $\hat{X}'_1$, $s_{\hat{x}'_1}$	$(2.01)s_{\hat{x}'_1}$	95% Limits	
					Lower	Upper
(1)	(2)	(3)	(4)	(5)	(6)	(7)
40	29.6	6.90	2.63	5.3	24.3	34.9
37	27.8	4.03	2.01	4.0	23.8	31.8
34	25.9	2.00	1.41	2.8	23.1	28.7
31	24.1	.81	.90	1.8	22.3	25.9
28	22.3	.46	.68	1.4	20.9	23.7
25	20.5	.95	.97	1.9	18.6	22.4
22	18.6	2.27	1.51	3.0	15.6	21.6
19	16.8	4.44	2.11	4.2	12.6	21.0
16	15.0	7.44	2.73	5.5	9.5	20.5

For 47 degrees of freedom, according to the table of Appendix E, the 95 percent confidence limits in units of t are +2.01 and −2.01. The predicted Reading Comprehension score from the sampling regression equation for the specified group is $\hat{X}_1 = (.61)(25) + 5.2 = 20.5$.

The confidence limits are, therefore, $20.5 - (2.01)(.974)$ and $20.5 + (2.01)(.974)$; or 18.5 and 22.5. These two values of X_1 are plotted in Fig. 9.1, with $X_2 = 25$ as the abscissa.

Similar points (lower and upper 95 percent confidence limits) for other specified values of X_2 permit plotting the two curved lines which form the limits of the band in Fig. 9.1. Some of these are shown in Table 9.2. In this table are included, as specified values of X_2, all the classmarks in the X_2 variable for the groups of the correlation chart in Fig. 8.4. In column 2 are shown the regression estimates, $\hat{X}'_1$, for the specified values in column 1. The pairs of figures in columns 1 and 2 are thus coordinates for points on the regression line of Fig. 9.1 The figures in column 3 are computed by means of equation 9.14, as illustrated above. The square root of these, appearing in column 4, are the standard errors of the respective predicted values, $\hat{X}'_1$. The width of the 95 percent band is determined in column 5. Finally, the limits appear in columns 6 and 7. The latter two columns of figures plotted respectively with the figures in column 1 as abscissas, form coordinates for points on the two curved lines in Fig. 9.1.

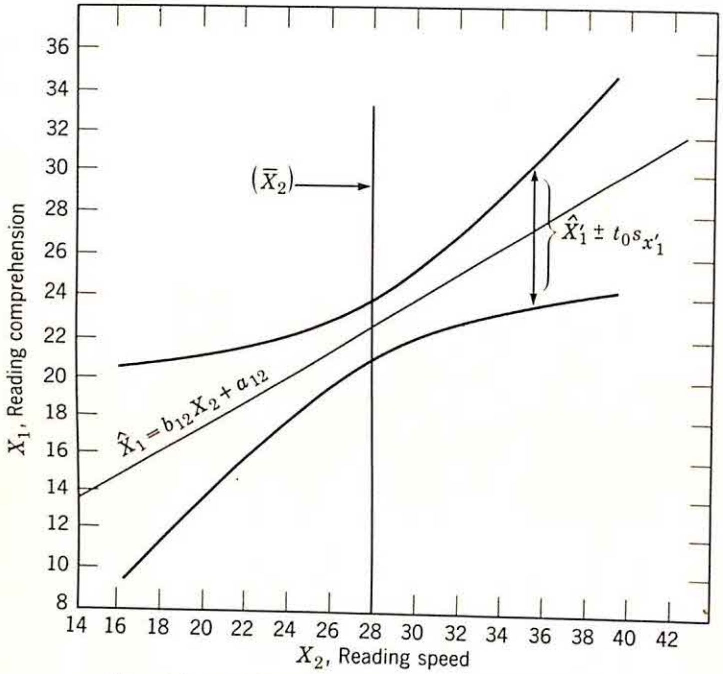


FIG. 9.1. Confidence bands for estimates by regression

9.10 THE TOTAL ERROR IN PREDICTING AN INDIVIDUAL SCORE FROM REGRESSION

It is important to note that the figures in column 4 of Table 9.2 deal *only* with the sampling distribution of a regression estimate, $\hat{X}_1$. These *standard errors* do *not*, therefore, tell us how much the regression estimate misses the actual X_1 score of an individual. This, as you will recall, is measured by the *residual variance*, $s^2_{1.2}$, or its square root, $s_{1.2}$, the so-called *standard error of estimate*. In considering how well, or how poorly, we may predict an individual X_1, given his score in X_2, we are really concerned with two questions involving error:

1. How far off is the sample regression estimate from the true regression estimate?

2. How far would the individual's actual score be from the true regression estimate?

The first of these is the *error of the regression estimate* and is measured by the variance, $s^2_{\hat{x}_1'}$, of equation 9.14. The second is the *residual variance*, $s^2_{1.2}$. The total error variance in predicting an individual's actual score is their sum. Hence, letting $s^2_{\text{Est}\,\hat{x}_1}$ represent this total variance, and X_2 a given score in the independent variable,

$$s^2_{\text{Est}\,\hat{x}_1} = s^2_{1.2} + s^2_{1.2}\left[\frac{1}{n} + \frac{(X_2 - \bar{X}_2)^2}{\Sigma x_2^2}\right]$$

or

$$s^2_{\text{Est}\,\hat{x}_1} = s^2_{1.2}\left[1 + \frac{1}{n} + \frac{(X_2 - \bar{X}_2)^2}{\Sigma x_2^2}\right] \tag{9.15}$$

The appropriate number of degrees of freedom is $n - 2$.

From the data of Table 9.2, we thus find, for example, that $s^2_{\text{Est}\,\hat{x}_1}$, *the total error variance* in predicting an individual Reading Comprehension score (X_1), given Reading Speed $X_2 = 37$, is $22.5 + 4.03 = 26.5$. Extracting the square root, $s_{\text{Est}\,\hat{x}_1} = 5.1$. For d.f. $= 47$, the 95 percent confidence limits for the *actual score*, X_1, for an individual for whom $X_2 = 37$, are thus $27.8 \pm (2.01)(5.1)$, or 17.5 and 38.1.

9.11 ERROR OF MEASUREMENT AND TEST RELIABILITY

The concepts of "random error" and correlation are basic to conventional theory of educational measurement. One method of defining the *reliability* of a test, for example, is in terms of the correlation of two theoretically "parallel" forms of the test given to a theoretical *population* of subjects. In this section we shall consider this definition and one of the several methods available for estimating error of measurement and test reliability.

We assume a test composed of items. The test score is the sum of item scores. We assume that an individual *actual score*, X_a, on the test consists of two components: the *true score* which the test would yield in the absence of error, and the error itself. This can be expressed by the relationship

$$X_a = X_t + e \tag{9.16}$$

where

$$X_a = \text{actual or obtained score}$$

$$X_t = (X_a - e) = \text{true score}$$

$$e = \text{error}$$

If we assume that the two components, X_t and e, are uncorrelated, the population variance of actual test scores is the sum of the variance of the two components:

$$\sigma_a^2 = \sigma_t^2 + \sigma_e^2 \tag{9.17}$$

where σ_a^2 is the variance of actual scores, σ_t^2 is the variance of *true scores*, and σ_e^2 is the *variance of errors of measurement*. The square root of the last variance, σ_e, is called the *standard error of measurement*. It is evident from equation 9.17 that the error component has the effect of inflating the variance of test scores. The larger the error, the greater the difference between the *actual* variance, σ_a^2, and the *true* variance, σ_t^2.

To simplify the analysis we assume also that in a large population of scores the mean of the errors is zero. Hence,

$$E(e) = 0$$

and

$$\begin{aligned} E(X_a) &= E(X_t + e) \\ &= E(X_t) + E(e) \\ &= E(X_t) \end{aligned}$$

Accordingly the population mean of *actual* (or *observed*) scores and the population mean of *true* scores are the same, μ_x. If we subtract this population mean μ_x from each side of equation 9.16, scores and components of scores in deviation form may be expressed as

$$x_a = x_t + e \tag{9.18}$$

The correlation of actual and true scores leads to one conception of error of measurement. This correlation may be expressed as

$$\rho_{at} = \frac{\Sigma x_a x_t}{N\sigma_a \sigma_t}$$

From equation 9.18 we may substitute $(x_t + e)$ for x_a, and from equation 9.17 we may substitute $\sqrt{\sigma_t^2 + \sigma_e^2}$ for σ_a, obtaining

$$\rho_{at} = \frac{\Sigma x_t^2 + \Sigma x_t e}{N\sigma_t \sqrt{\sigma_t^2 + \sigma_e^2}}$$

Since X_t and e are assumed uncorrelated, $\Sigma x_t e = 0$, and

$$\rho_{at} = \frac{\sigma_t^2}{\sigma_t \sqrt{\sigma_t^2 + \sigma_e^2}}$$

$$= \frac{\sigma_t}{\sqrt{\sigma_t^2 + \sigma_e^2}} \tag{9.19}$$

This correlation, termed the *index of reliability*, is the correlation of actual observed scores on a test with true scores of the test.

Squaring equation 9.19, we obtain

$$\begin{aligned} \rho_{at}^2 &= \sigma_t^2/(\sigma_t^2 + \sigma_e^2) \\ &= \sigma_t^2/\sigma_a^2 \end{aligned} \tag{9.20}$$

That is, the *square* of the correlation between actual and true scores is the proportion of true variance in the variance of actual scores. This may also be written $\rho_{at}^2 = (\sigma_a^2 - \sigma_e^2)/\sigma_a^2$.

Solving for σ_e^2, we have

$$\sigma_e^2 = \sigma_a^2(1 - \rho_{at}^2) \tag{9.21}$$

Comparing this with equation 8.27, we note that σ_e, the *standard error of measurement*, is the same as the *standard error of estimate* of the regression of X_a on X_t.

The foregoing relationships do not permit the determination of σ_e^2, σ_t^2, or ρ_{at}. With only one administration of a test we do not know either X_t or e, the two components of our observed measures. However, if we have repeated measures of the same variable, or if we have two parallel tests, we may correlate them and estimate the variance components σ_e^2 and σ_t^2. What we are leading up to is the very simple notion of reliability as "self-correlation." However, we can only imagine this in most educational measurements because it is possible to administer only one test at a time; and once a test has been administered changes take place in subjects (partly as a consequence of the learning that takes place at the time of administering the test), so that a test will measure something different when repeated. It is not realistic to assume two identical measures of the same thing at the same time. That is why the development is somewhat more logical if we think in terms of two parallel forms of a test.

Parallel tests may be defined in various ways. For our purposes it will be sufficient if we consider two tests parallel when, for any individual, the two *true* scores X_{t1} and X_{t2} are identical; if the error components of the two tests are uncorrelated; and if their error variances, σ_{e1}^2 and σ_{e2}^2, are equal. Then the variances of the actual scores, σ_1^2 and σ_2^2, will be equal.

From equation 9.18, deviation scores for the two forms of the test may be written as

$$x_1 = x_t + e_1$$

and (9.22)

$$x_2 = x_t + e_2$$

The correlation of the two forms may be written

$$\rho_{12} = \frac{\Sigma x_1 x_2}{N\sigma_1\sigma_2}$$

Substituting from equation 9.22, the numerator of this correlation coefficient may be expressed as

$$\Sigma(x_t + e_1)(x_t + e_2) = \Sigma x_t^2 + \Sigma x_t e_1 + \Sigma x_t e_2 + \Sigma e_1 e_2$$

If errors are uncorrelated with true scores and if errors of one test are uncorrelated with errors of the other, the last three sums of products are zero and the above expression for the numerator of ρ_{12} reduces to Σx_t^2. Hence

$$\rho_{12} = \frac{\Sigma x_t^2}{N\sigma_1\sigma_2} = \frac{\sigma_t^2}{\sigma_1\sigma_2}$$

By our definition of parallel tests, $\sigma_1 = \sigma_2 = \sigma_a$. Therefore,

$$\rho_{12} = \frac{\sigma_t^2}{\sigma_a^2} = 1 - \frac{\sigma_e^2}{\sigma_a^2} \tag{9.23}$$

This is identical to the theoretical value given in equation 9.20 of ρ_{at}^2, the square of the correlation between *observed* and *true* scores in a test.

The correlation, ρ_{12}, of two measures of the same thing (as represented by the correlation of two parallel forms of a test) is termed the *coefficient of reliability* or simply the *reliability coefficient*. It is usually defined in measurement literature as in equation 9.23, the proportion of variance in observed test scores which is due to true variance among individuals.

Noting from equations 9.20 and 9.23 that $\rho_{12} = \rho^2_{at}$, we substitute the reliability coefficient in equation 9.21. This gives us an expression for the variance of errors of measurement,

$$\sigma^2_e = \sigma^2_a(1 - \rho_{12}) \tag{9.24}$$

The square root of this is σ_e, the *standard error of measurement.*

In practice ρ_{12} is estimated by r_{12}, a correlation based upon a sample of the population. Since r_{12} is a straightforward correlation of two variables X_1 and X_2, the methods of other sections in this chapter may be used to establish confidence limits for ρ_{12} and for testing hypotheses concerning ρ_{12}. This is an important step in the evaluation of test results. The reliability coefficient and error of measurement are functionally related and, as has been emphasized, the *error of measurement* is interpretable as *error of estimate.* Hence it is a measure of residual variation in the correlation model. As such it is not a measure of *sampling error.* Rather it should be remembered that estimates of reliability and of error of measurement derived from samples are themselves subject to sampling errors.

Also of importance in the interpretation of reliability (and its complement, error of measurement) is the kind of error taken into account by the experimental conditions under which the data were obtained. A reliability coefficient from two forms of a test administered at two different times, for instance, will reflect a combination of two types of measurement error: (*a*) failure of the two forms to be *equivalent* (that is, measure the same thing) and (*b*) absence of *stability* (that is, failure of the scores for whatever the test measures to remain the same from time to time.)

True score may be defined as the expected value of an infinite number of repeated measures of the same thing on a single individual. This conception of a *true* score will be used in another approach to reliability in Section 12.10. Equations 9.23 and 9.24 may be derived from that conception of true score, but such a derivation is somewhat more involved.

9.12 THE SPEARMAN-BROWN FORMULA

Since it is not always possible to have two parallel forms of a test, methods have been developed for estimating reliability from a single administration of a test. One such method splits the items of a test into two or more parts and correlates the parts, thus estimating the reliability of the parts. To estimate the reliability of the entire test it is necessary to know the relationship of reliability of parts to the reliability of the total.

In this section we examine the simplest case, that of estimating reliability of a test from the correlation between two parts, each of which consists of half of the items of the total test. This correlation estimates the reliability of a test half the length of the total test. We look upon the two halves of the test as two parallel tests. Then the variance of their sum is the sum of the variances of their actual scores plus twice the covariance (equation 11.4). That is

$$\sigma^2_{(1+2)} = \sigma^2_1 + \sigma^2_2 + 2\sigma_1\sigma_2\rho_{12}$$

But since the two parts are considered parallel tests, by definition, $\sigma^2_1 = \sigma^2_2 = \sigma^2_a$. Therefore

$$\begin{aligned}\sigma^2_{(1+2)} &= 2\sigma^2_a + 2\sigma^2_a\rho_{12}\\ &= 2\sigma^2_a(1 + \rho_{12}) \qquad (9.25)\end{aligned}$$

The variance of the sum of *true* scores of the two parallel parts is similarly the sum of the two true score variances plus twice the covariance:

$$\sigma^2_{(t_1+t_2)} = \sigma^2_{t_1} + \sigma^2_{t_2} + 2\sigma_{t_1}\sigma_{t_2}\rho_{t_1t_2} \qquad (9.26)$$

By the definition of parallel tests, the true scores are identical. Therefore their variances are equal and the correlation between them is 1. Equation 9.26 thus reduces to

$$\sigma^2_{(t_1+t_2)} = 4\sigma^2_t \qquad (9.27)$$

Substituting in equation 9.23, the coefficient of reliability of the whole test is

$$\rho_{(1+2)} = \frac{\sigma^2_{(t_1+t_2)}}{\sigma^2_{(1+2)}} = \frac{4\sigma^2_t}{2\sigma^2_a(1 + \rho_{12})} = \frac{2\rho_{12}}{1 + \rho_{12}} \qquad (9.28)$$

This is known as the Spearman-Brown formula for finding the reliability of the *total* test from the reliability of *half* the test, or for finding the reliability of a test double the length of a given test. It is most commonly used with the correlation of "split-halves" of a test to estimate the reliability of the total test.

The Spearman-Brown formula may be extended by the derivation of the reliability of a composite of k parallel forms of a test. The result is the following formula for finding the reliability of a test if increased in length k times:

$$\rho_c = \frac{k\rho_{12}}{1 + (k - 1)\rho_{12}} \qquad (9.29)$$

where ρ_{12} is the reliability coefficient of the test and ρ_c is the reliability coefficient of a test k times in length.

In this book we consider only a few simple examples of statistics of tests and measurements. The treatment in this chapter is thus but an introduction to measurement theory. The student planning extended use of educational tests or the development of such tests should make a special study of that subject and consult references 4 and 5.

EXERCISES

1. What are the 95 percent confidence limits of ρ for the following?

$$(a)\ n = 12;\quad r = .75 \qquad (c)\ n = 100;\quad r = -.14$$
$$(b)\ n = 50;\quad r = .35 \qquad (d)\ n = 200;\quad r = .85$$

Which of the above correlations are significant at the .05 level?

2. Disregarding the grouping of the data in Appendix G, and assuming random sampling from a common population, find the 95 percent confidence limits for the correlation of X and Y based upon 159 observations.

3. Suppose a pair of dice is rolled twelve times and the points for each roll recorded in order so that the odd-numbered rolls are in the first column and the even-numbered rolls in the second. Let X_1 be the odd-numbered rolls and X_2 the even-numbered rolls. The first roll is paired with the second, the third with the fourth, and so on so that we may compute a correlation between the six pairs of values in X_1 and X_2. What is the expected value of r? As a class exercise let each student conduct the above experiment several times so that for the entire class there will be 50 or 100 sample correlation coefficients. Make a frequency distribution of the correlation coefficients observed in these 50 or 100 experiments. Plot on probability paper to see how close the distribution is to the normal. Compute t from formula 9.2 for each experiment. If each experiment had been used to test the hypothesis of $\rho = 0$, how many times and what proportion of the time would it have been rejected at the .05 level? Check the theory against the data.

4. From the information given in Ex. 16 of Chapter 8 test at the .05 level the significance of the correlation between current expenditures per pupil and property valuation per pupil of school districts.

5. Using the results of Ex. 3 of Chapter 8, test the hypothesis that the correlation between the California test and the writing skills test is zero. Use two methods, equations 9.1 and 9.2, and compare results, explaining such differences in results as you find.

6. From the results of Ex. 3 of Chapter 8 test the hypothesis,

$$H : \rho_{12} = .75.$$

Why is it necessary to use a different technique in testing this hypothesis from the one used in testing the hypothesis of Ex. 5 above?

7. From the results of Ex. 3 of Chapter 8 test the hypothesis that the regression coefficient for predicting scores on the Writing Skills Test from the California Test of Mental Maturity is zero. Compare this result with that of Ex. 5 above, and explain. Is the regression coefficient for the prediction of California test score from writing skills test score significant?

8. Using the data from Ex. 3 of Chapter 8 and the methods of Section 9.9, prepare a diagram similar to Fig. 9.1, showing the regression line for predicting the California test on the writing skills test and 95 percent confidence limits for such predictions.

9. From results of the previous exercise compute the total error in predicting a California test score from a writing test score of 21? of 42? of 54?

10. How would you estimate the distribution of all possible correlations of ten pairs of scores from Appendix A? Would you expect this to be a symmetrical distribution, a negatively skewed one, or a positively skewed one? Would the variance of this distribution be less than or greater than that of correlations from samples of size 20?

11. Select ten random pairs of digits from Appendix B. Let the first digit of each pair be X_1, the second digit X_2. Compute r and $s_{1.2}$. What is the expected value of r and $s_{1.2}$? Check this by averaging results of several repetitions of this experiment by all members of the class.

12. What error is measured by the *standard error of estimate*? Is the standard error of estimate a statistic or a parameter? In what respects might we consider the term "standard error of estimate" misleading? What other term might be used to avoid this difficulty?

13. *Random* errors of measurement are present in both X_1 and X_2.

(*a*) What effect has the presence of this error upon the covariance of X_1 and X_2?

(*b*) What effect on the variance of X_1 and X_2?

(*c*) What effect on the correlation between X_1 and X_2?

14. Two measures contain *systematic error*, that is, errors which produce uniformly high or uniformly low scores for all subjects. What effect has this error upon their correlation?

15. The correlation between two parallel forms of a 75-item test was found to be .78.

(*a*) What is the coefficient of reliability for the test?

(*b*) What proportion of the total variance in the test is due to error?

(*c*) What is the reliability of a score consisting of adding together the scores of both forms of the test administered at one time?

(*d*) If the observed correlation coefficient between the two forms is based upon a sample of 150 subjects, what are the 95 percent confidence limits for the reliability coefficient?

16. In a study by Cornell, McLure, Miller, and Wochner, "Financing Education in Efficient School Districts," costs of transportation in theoretically "efficient districts" were estimated from maps showing the location of dwellings for all such districts in Illinois. Acceptable estimates of costs may be determined from locations of actual pupils and the application of standards on the routing of school buses. Dot maps of pupils were not available for the theoretical districts. It was possible, however, to obtain "pupil dot maps" for a sample of 31 "unit districts" organized very much along lines of the proposed "efficient" districts. For these 31 districts standard routing practices were applied and "criterion costs" of transportation, X_1, were determined. Maps showing the location of dwellings were available for all districts in the state. From these a measure was derived which is the number of dwellings one mile or more from the school center weighted by distance from school in miles. The assumptions were:

(1) Cost is a function of the geographical distribution of pupils.

(2) The geographical distribution of pupils is a function of the geographical distribution of dwellings.

The dwelling measure, X_2, was thus used to predict transportation costs X_1.

The results of the sample study are as follows:

X_1 = total estimated annual cost of transportation

X_2 = dwelling-miles

$n = 31$

$$\Sigma X_1 = 471{,}700 \qquad \Sigma X_2 = 99{,}900$$

$$\Sigma X_1^2 = 8{,}617{,}150{,}000 \qquad \Sigma X_2^2 = 459{,}010{,}000$$

$$\Sigma X_1 X_2 = 1{,}933{,}210{,}000$$

$$s_1^2 = 47{,}990{,}064 \qquad s_2^2 = 4{,}569{,}140$$

$$s_{1.2}^2 = 6{,}711{,}420$$

$$\hat{X}_1 = 3.0138 X_2 + 5{,}504$$

(*a*) What is the correlation between X_1 and X_2? Would you expect this to suggest that X_1 may be predicted from X_2?

(*b*) What are the 95 percent confidence limits for a_{12} and b_{12}?

(*c*) Find the standard error of the regression estimate of cost for a group of districts whose dwelling-mile measure is 5,000. Compute the cost estimate for these schools. What is the ratio of the standard error to the estimate (the coefficient of variation)?

(*d*) What are the 95 percent confidence limits for estimating an individual district's annual cost, X_1, from the regression equation, when X_2 is given as 5,000?

REFERENCES

1. Dixon, Wilfrid J., and Frank J. Massey, *Introduction to Statistical Analysis*, New York, McGraw-Hill Book Co., 1951, Chapter 11.
2. Edwards, Allen L., *Experimental Design in Psychological Research*, New York, Rinehart and Co., 1951, Chapter 7.
3. Guilford, Joy P., *Fundamental Statistics in Psychology and Education*, Second Ed., New York, McGraw-Hill Book Co., 1950, Chapters 17 and 19.
4. Guilford, Joy P., *Psychometric Methods*, Second Ed., New York, McGraw-Hill Book Co., 1954, Chapters 13 and 14.
5. Gulliksen, Harold, *Theory of Mental Tests*, New York, John Wiley and Sons, 1950.
6. Mood, Alexander M., *Introduction to the Theory of Statistics*, New York, McGraw-Hill Book Co., 1950, pp. 289-99.
7. Walker, Helen M., and Joseph Lev, *Statistical Inference*, New York, Henry Holt and Co., 1953, Chapter 10.

CHAPTER 10

Chi Square and Enumeration

In Chapter 2 a distinction was made between counting and enumeration, on the one hand, and continuous measurement on the other. Of particular use in the statistical treatment of enumeration data (that is, frequencies) is the chi-square (χ^2) distribution. This distribution is of considerable theoretical significance and is related mathematically to other functions, such as the t distribution, which are important in statistical inference. Its chief value in practice is in a large variety of problems involving the comparison of *observed* and *theoretical* frequencies.

We have used the normal distribution and the t distribution for purposes of testing hypotheses about parameters. We saw that assumptions had to be made about the distribution tested in order that these tests would be valid. There are several statistical tests which may be used for hypotheses concerning distributions (not parameters), and which do not require specification of the form of the distribution involved. The statistics used for such tests are called *nonparametric*. χ^2 is one such statistic.

10.1 COMPARING OBSERVED AND THEORETICAL FREQUENCIES

In general, observed and theoretical frequencies may be compared by calculating the statistic χ^2, defined by

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_0 - f_t)^2}{f_t} \tag{10.1}$$

where k is the number of categories into which frequencies are grouped, and f_0 and f_t are the *observed* and *theoretical* frequencies, respectively.

We shall illustrate the application of equation 10.1 by an experiment concerning the Table of Random Numbers in Appendix B. If the table is truly a table of random numbers, we would expect, in the long run from a sample from this table, equal frequencies of the digits 0 to 9 inclusive. That is, in a sample of 200 numbers, 20 theoretical zeros would be expected, 20 ones, 20 twos, and so on. The results of tabulating the actual frequencies of 200 numbers systematically drawn from the table appear in column 1 of Table 10.1. Here we naturally see some discrepancy between the actual (observed) frequencies and the expected (theoretical) of 20 for each digit (as shown in column 2) under the hypothesis of "equal probability" of the numbers.

TABLE 10.1

Chi-Square Test of 200 Random Numbers

Number	f_0	f_t	$f_0 - f_t$	$(f_0 - f_t)^2$	$\frac{(f_0 - f_t)^2}{f_t}$
	(1)	(2)	(3)	(4)	(5)
0	20	20	0	0	0.00
1	16	20	−4	16	0.80
2	15	20	−5	25	1.25
3	23	20	3	9	0.45
4	28	20	8	64	3.20
5	19	20	−1	1	0.05
6	21	20	1	1	0.05
7	14	20	−6	36	1.80
8	19	20	−1	1	0.05
9	25	20	5	25	1.25
	200	200	0	—	8.90

It may be noted that χ^2 as computed by equation 10.1 is a measure of discrepancy of disagreement between the observed and the theoretical frequencies. In order to compute if we first find the difference between each pair of frequencies in column 3, then square the result in column 4, and finally express the square of the difference as a ratio to f_t as in column 5. We note that the figures in column 5 are, thus, *relative* squared "discrepancies" between the observed and the theoretical frequencies.

The sum over all ten sets of figures in column 5 yields $\chi^2 = 8.90$. Note that the frequency for the number zero agrees perfectly with expectation, and zero discrepancy is recorded in columns 3, 4, and 5. Clearly, if there had been perfect agreement between the observed and the theoretical frequencies, the figures in columns 3, 4, and 5 would all be zero and χ^2

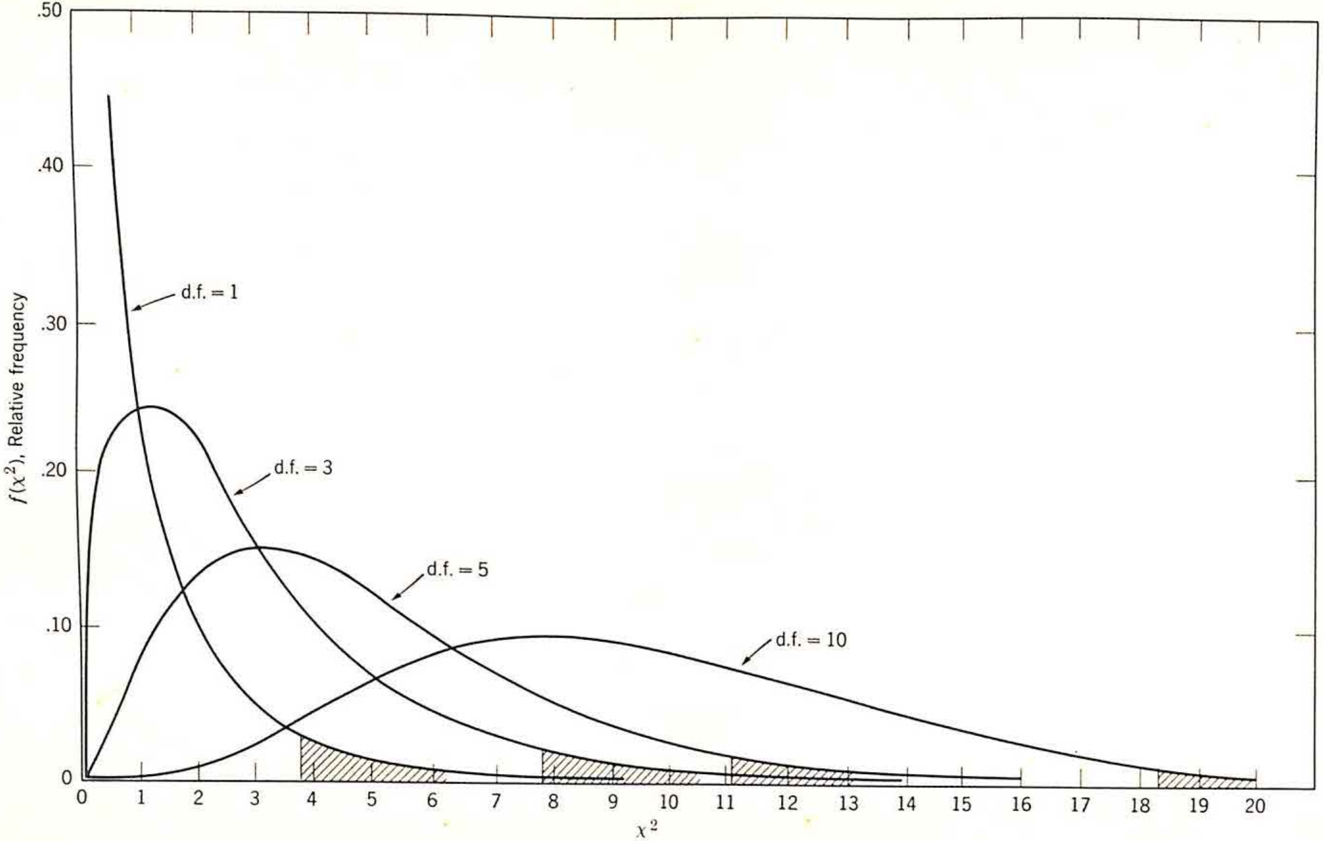


FIG. 10.1. Chi square distribution and 5 percent critical regions, various degrees of freedom.

would be zero. Obviously, the *greater the disagreement* of observed and theoretical frequencies, the greater the value of χ^2.

10.2 THE DISTRIBUTION OF CHI SQUARE

Now we might think of many such chi squares computed for each of a large number of random samples of 200 drawn from a population containing equal numbers of the 10 digits. These chi squares could be arranged in a relative frequency distribution which would enable us to tell the proportions (or probabilities) of various values of χ^2 which should be expected. This would be a guide for all comparisons, such as those in Table 10.1, involving 10 pairs of frequencies. However, it would not serve for comparisons of more than or fewer than 10 pairs. If we computed χ^2 for only 4 pairs, we would expect a smaller value; for 25 pairs a larger value. What is needed, therefore, is a family of distribution functions, one for each possible number of "independent" comparisons.

At this point we will make further use of the term "degrees of freedom." In our example, we have assumed that each observation falls into one, and only one, of the classes 0–9, and that the ten frequencies sum to 200, that is, $\Sigma f_0 = 200$. Similarly, $\Sigma f_t = 200$. Hence, after we had made any 9 of the 10 comparisons, the remaining comparison would be determined by the amount needed to bring the total of f_0 and f_t to 200. There is, thus, a restriction imposed upon the f_0, and we consider the degrees of freedom by which the observed may vary from the theoretical to be $k - 1$, or 9.

For such problems, a continuous function, known as the χ^2 distribution, may be used. It is really a family of distributions, like the t distribution, one for each number of degrees of freedom. The distribution of χ^2 for various numbers of degrees of freedom is shown graphically in Fig. 10.1.

It is not important in elementary statistics to know formulas for the many distribution functions used in statistical work, but to use tables intelligently it is important to realize that such functions exist, that they may be graphed, and that tabled values may be computed from them. For our purposes, it is of passing interest to note that ordinates for χ^2 curves, like those in Fig. 10.1, may be computed from the function

$$f(\chi^2) = \frac{1}{[(\nu - 2)/2]!\, 2^{\nu/2}} (\chi^2)^{\frac{\nu-2}{2}} e^{-\chi^2/2} \tag{10.2}$$

in which ν is the number of degrees of freedom.

Tables have been developed which show for various degrees of freedom the values of χ^2 which will be exceeded with given probabilities. Such a table appears as Appendix H. The P value is the probability that, on random sampling, we should get a value of χ^2 *as great as, or greater than*, the value of χ^2 shown, for various degrees of freedom. The values of χ^2 for $P = .05$ and $P = .01$ are critical values for statistical tests at the 5 percent and 1 percent levels, respectively—the levels most commonly used in testing statistical hypotheses. The table in Appendix H gives χ^2 values which determine the critical regions for other levels of risk. The shaded areas of Fig. 10.1 are 5 percent critical regions ($\alpha = .05$) for the four χ^2 curves shown.

It will be noted that the distributions in Fig. 10.1 are continuous, that is, smooth curves. The actual distributions of χ^2 as computed from equation 10.1 are discrete. However, the mathematical function, $f(\chi^2)$, is a close approximation to the distribution of equation 10.1 and is used in much the same way that we use the normal distribution as an approximation to the binomial. It is important to remember that there are thus really two chi squares: (1) the χ^2 distribution function, equation 10.2, and (2) the computed value of χ^2 from equation 10.1 which is distributed *approximately* as the distribution function. Understanding this will help avoid common mistakes in the use of χ^2.

10.3 TESTING HYPOTHESES CONCERNING FREQUENCIES

Knowing the approximate distribution of the χ^2 defined by equation 10.1, we may now test the hypothesis $H : f_0 = f_t$, that the sample of 200 numbers in Table 10.1 is drawn from a population with equal frequencies for the 10 digits. If a computed χ^2 has a very low probability of occurrence, we reject the hypothesis. For instance, we find by referring to Appendix H that χ^2 for 9 d.f. is 16.9 at the 5 percent level. This means that an observed χ^2 of as much as 16.9 may be expected to occur in 5 percent of the cases purely by chance. That is, $P(\chi^2 \geq 16.9) = .05$. Our computed χ^2 of 8.90 is therefore not great enough to justify rejecting the hypothesis at the 5 percent level. Further reference to the χ^2 table shows that the probability of χ^2 of 8.90 is between .50 and .30 for 9 d.f.

It is pointed out that the rejection regions that we used in the table of Appendix H, and as shown in Fig. 10.1, will not enable us to reject an hypothesis on the basis of *unusually low* values of χ^2. We would, of course, be as ready to conclude that the data were *not* a random sample on the hypothesis $H : f_0 = f_t$ if χ^2 turned out to be unusually small as we would if χ^2 turned out to be unusually large.

As a further example of the use of χ^2, a sample of 50 trainees who had attended public vocational training courses were grouped, according to years of school completed, as follows:

Years of Schooling	f_0
13 and over	6
9–12	32
5–8	7
0–4	5
	$n = 50$

We wish to test the hypothesis that this is a random sample with reference to schooling from the general population 25 years of age and over. On the basis of census statistics we find that the schooling of the population 25 years of age and over is distributed as follows:

Years of Schooling	*Percent of Population*
13 and over	10.1
9–12	29.5
5–8	46.7
0–4	13.7
	100.0

Fifty cases distributed according to the above percentages would give us theoretical frequencies of 5, 15, 23, and 7, as compared respectively with the *observed* frequencies of 6, 32, 7, and 5. Applying formula 10.1, $\chi^2 = (6 - 5)^2/5 + (32 - 15)^2/15 + (7 - 23)^2/23 + (5 - 7)^2/7$, or 31. We note in Appendix H that for 3 d.f. the .01 level for χ^2 is 11.34. The probability of a χ^2 as great as 31 is hence considerably less than 1 percent. Therefore we reject the hypothesis at the 1 percent level.

10.4 TESTING GOODNESS OF FIT

A special type of application of the foregoing procedure called "testing goodness of fit" concerns the comparison of a set of observed frequencies with a set of theoretical frequencies derived from fitting some type of curve or distribution function to the observed data. In Chapter 6 we fitted the normal curve to the distribution of scores of 500 naval recruits. The normal curve in Fig. 6.8 is seen to follow fairly closely the contour of the histogram of the actual 500 scores. The actual or observed frequencies are shown in Table 6.2 with theoretical frequencies based on the normal distribution. We shall proceed with formula 10.1 to compute χ^2 to test the hypothesis that the 500 scores are a sample from a normal population.

In this problem there are two additional steps in our procedure. The first concerns the small frequencies at the ends of the distribution. These we shall eliminate by grouping, because χ^2 computed from equation 10.1 is only *approximated* by the function tabled in Appendix H, and the approximation is better if no expected frequency is too small (say, less than 5).

The second concerns the degrees of freedom. In this problem there are *three* restrictions on the number of *independent* comparisons which may be made. That is, d.f. $= k - 3$. A review of the steps discussed in Section 6.6 shows that the determination of the theoretical frequencies required that we specify three constants for the theoretical normal distribution. These were N, μ, and σ. That is, in fitting the normal curve we use the number of cases, the mean, and the standard deviation of the data. Each of these imposes a restriction on the theoretical frequencies, which may therefore differ from the observed frequencies in three fewer independent respects than the number of categories into which the frequency distribution is subdivided.

TABLE 10.2

CHI-SQUARE TEST OF NORMALITY OF DISTRIBUTION

Score	f_0	f_t	$f_0 - f_t$	$(f_0 - f_t)^2$	$\frac{(f_0 - f_t)^2}{f_t}$
85 and over	7	5.5	1.5	2.25	0.41
80–84	11	7.4	3.6	12.96	1.75
75–79	13	14.4	−1.4	1.96	0.14
70–74	18	24.9	−6.9	47.61	1.91
65–69	37	38.4	−1.4	1.96	0.05
60–64	50	52.4	−2.4	5.76	0.11
55–59	72	63.6	8.4	70.56	1.11
50–54	70	68.6	1.4	1.96	0.03
45–49	56	65.5	−9.5	90.25	1.38
40–44	58	55.9	2.1	4.41	0.08
35–39	50	42.2	7.8	60.84	1.44
30–34	30	28.3	1.7	2.89	0.10
25–29	13	16.9	−3.9	15.21	0.90
20–24	9	9.0	0	0.00	0.00
19 or less	6	7.0	−1.0	1.00	0.14
Total	500	500.0	0.0	——	9.55

The χ^2 test of goodness of fit of the normal distribution in Table 6.2 is computed in Table 10.2. Here we note that the highest class is 85 and over, because the three top theoretical (normal) frequencies in Table 6.2 were each less than 5. The rule mentioned above was followed in grouping these three to yield the top category with an observed frequency of 7 and a theoretical frequency of 5.5. Similarly, in Table 6.2 the last three categories had normal theoretical frequencies of less than 5. These are grouped in Table 10.2 into the single category of 19 or less, with an observed frequency of 6 and a theoretical frequency of 7.0. The remainder of Table 10.2 is a straightforward application of equation 10.1. In this table, after the combining of groups with small theoretical frequencies, 15 categories remain. Therefore, d.f. $= 15 - 3 = 12$.

Reference to the table in Appendix H shows that χ^2 must be at least 21 for us to reject the hypothesis at the 5 percent level. The observed χ^2 is well within the acceptance region and, therefore, the hypothesis of normality is tenable for the distribution of the 500 scores.

10.5 VARIATIONS IN COMPUTATION PROCEDURE

Formula 10.1 may be expressed in various equivalent forms which simplify computational work for some types of problems. One version of formula 10.1, for instance, is particularly useful where theoretical frequencies are fractional. It avoids the necessity of squaring decimal fractions and dividing by decimal fractions as is required in equation 10.1. This equivalent form is as follows:

$$\chi^2 = \sum_{i=1}^{k} (f_0^2/f_t) - N \tag{10.3}$$

where $N = \Sigma f_0 = \Sigma f_t$.

In computing χ^2 to test the hypothesis concerning the distribution of years of schooling for the sample of 50 trainees, we compute

$$\begin{aligned}\chi^2 &= (36/5 + 1024/15 + 49/23 + 25/7) - 50\\ &= 81.17 - 50.00 = 31.17\end{aligned}$$

This verifies the numerical equivalence of computation from equation 10.3 with our previous calculation of χ^2 by equation 10.1.

Some of the many other special methods of computing χ^2 will be presented in connection with special problems discussed in later sections of this chapter.

A modification of the procedure for determining critical levels of χ^2 is needed in conjunction with Appendix H for occasions when the number of degrees of freedom exceeds 30. We note that values tabled in Appendix H are for distributions of χ^2 for 30 d.f. or fewer. There are two reasons for this. One is that few problems involve more than 30 d.f. The chief reason, however, is that the χ^2 distribution for large degrees of freedom may be adequately approximated by the normal distribution. A glance at Fig. 10.1 suggests this, even with d.f. = 10. The shape of the χ^2 distribution for d.f. = 10 is more symmetrical, and appears more like the bell-shaped normal distribution, than distributions for smaller degrees of freedom.

The function

$$z = \sqrt{2\chi^2} - \sqrt{2(\text{d.f.}) - 1} \tag{10.4}$$

is distributed approximately as the normal deviate, z, and has mean 0 and unit standard deviation, when d.f. is larger than 30.

For instance, if we obtain $\chi^2 = 49.5$ with d.f. = 39, we may compute

$$\begin{aligned} z &= \sqrt{(2)(49.5)} - \sqrt{(2)(39)} - 1 \\ &= 9.95 - 8.77 \\ &= 1.18 \end{aligned}$$

Reference to the cumulative normal function in Appendix D shows us that $P(z > 1.18) = .1190$. In other words, with 39 d.f. a χ^2 of 49.5 or more should be expected about 12 percent of the time. Therefore, the observed χ^2 would not be considered significant. Remembering that the critical region we use for χ^2 is the right-hand tail of the distribution, all values of z, computed from equation 10.4, greater than $z_{.95} = 1.64$ would be in the 5 percent critical region, and values of z greater than $z_{.99} = 2.33$ would be in the 1 percent critical region.

10.6 TESTS OF INDEPENDENCE AND HOMOGENEITY

In the foregoing examples illustrating applications of χ^2, theoretical frequencies have been derived by applying *a priori* probabilities to the total of all k classes. The *a priori* probabilities assigned to each of the digits in Table 10.1 was 1/10. The *a priori* probabilities for the sample of 50 trainees was derived from the percentage distribution for the total popu'ation.

In a large class of problems involving the use of χ^2, theoretical frequencies are derived from the observed experimental data themselves. In the parentheses in Table 10.3 are shown theoretical frequencies derived in this manner. This is a table showing the distribution of a sample of 125 persons interviewed in an opinion poll in a school survey. The 125 respondents are classified in a two-way table. One set of categories is for responses on a question regarding their approval or disapproval of practices which had been recently introduced in the school system. By means of information concerning the occupation of the head of household and observation of the material characteristics of the home, respondents were also classified according to economic class. Thus the body of Table 10.3 consists of three rows and three columns. The observed frequencies may be read in some such manner as this: 14 of 52 persons classified in the upper-middle economic class indicated approval of the school program, and 14 of the 48 persons who indicated approval of the school program were classified in the upper-middle economic class. Our objective is to learn whether there is any relationship between responses to the question and the economic level of the respondent. Is a person in one economic class more likely to approve than persons in another economic class? In short, are the proportions approving, disapproving, and undecided the same in each economic class?

TABLE 10.3

DISTRIBUTION OF SAMPLE OF RESPONDENTS IN SAMPLE SURVEY

Economic Class	Attitude toward School System			Total
	Approve	Disapprove	Undecided	
Upper middle	14 (19.97)	18 (14.56)	20 (17.47)	52
Middle	22 (16.90)	10 (12.32)	12 (14.78)	44
Lower	12 (11.14)	7 (8.12)	10 (9.74)	29
Total	48	35	42	125

To use equation 10.1 we need the theoretical frequencies in the body of the table for the hypothesis that each of the three classes responds in the same proportion and that each of the response categories is distributed proportionally according to economic class. We would then use χ^2 to test whether the observed frequencies differed from the theoretical by such an amount that the variation could not reasonably be attributed to chance.

Such a significance test is called a test of *independence*. The theoretical values must in this sense be those values to be expected on the hypothesis of proportionality in the table or, to put it another way, the hypothesis of *no relationship* between the two criteria of classification.

The theoretical frequencies are derived from the marginal totals in the table. In the last column of the table we see that 52 of the total of 125 respondents, or 41.6 percent, were in the upper-middle class. Since there was a total of 48 persons responding "approve" there would need to be $(48)(.416) = 19.97$ upper-middle class persons responding "approve" if the two attributes—attitude towards school and economic class—are independent. By similar reasoning we can compute each of the theoretical frequencies shown within parentheses in Table 10.3.

In actual computation we simply multiply, for any cell of the table, the corresponding row total by the column total and divide the result by the grand total. That is, $f_{t_i} = n_{r_i}n_{c_i}/N$, where f_{t_i} is a theoretical frequency for the ith cell, n_{c_i} and n_{r_i} are respectively the column and row totals for the column and row in which the ith cell is located, and N is the grand total. By means of either equation 10.1 or 10.3 we can compute χ^2, which is found to be 5.69.

In an $r \times c$ table such as this, only $(r - 1)(c - 1)$ cells can be filled in arbitrarily. In each row and each column the theoretical frequencies must add to the corresponding row and column totals. After computing any two of the theoretical frequencies for any row in Table 10.3, the remaining one must make up the remainder of the row total. The same is true of columns. In this table there are thus only four frequencies in the body of the table which may be assigned independently. We therefore have d.f. $= (r - 1)(c - 1) = (3 - 1)(3 - 1) = (2)(2) = 4$.

From the table in Appendix H we see that a χ^2 would have to be more than 9.48 before we would reject at the 5 percent level. Therefore, we *do not* reject the hypothesis of independence of the two factors of classification represented in Table 10.3

In testing *independence*, it should be noted that χ^2 is a measure of *absence* of independence, that is, the lack of agreement between actual frequencies and theoretical frequencies which would obtain if there were proportionality. This absence of independence is sometimes called *interaction*. If we had *rejected* the hypothesis of independence between the two schemes of classification in Table 10.3, we would simultaneously have established that there was a significant interaction between the two variables of classification. This would mean that response on the attitude question depended upon economic class, or that economic class was related to response to the question.

In the previous illustration we were dealing with a *single* sample of 125 individuals classified in two ways. Sometimes a two-way classification of frequencies is formed by classifying several *separate* samples on some scheme of classification. Table 10.4 is such a table of frequencies.

This table is made up of four *separate* samples from the four separate high schools of Appendix G.

TABLE 10.4

NUMBER OF BOYS AND GIRLS IN SAMPLES OF STUDENTS IN FOUR HIGH SCHOOLS
(From Appendix G)

High School	Number of		
	Boys	Girls	Total
A	22	23	45
B	21	14	35
C	17	18	35
D	27	17	44
Total	87	72	159

In the example of Table 10.3 we were interested in examining the homogeneity of a single sample on one classification relative to the other. In the example of Table 10.4 our interest is in the comparison of the sex composition of four separate samples. The difference between the two types of problem is in the objective and in the source of data. There is no difference in the application of χ^2. For Table 10.4 our hypothesis is $H : p_i = p$, that the proportion of boys, p_i, in the population supplying the ith high school, is the same for all four schools and equal to p, the proportion of boys in the *total* population. The hypothesis is proportionality in the table just as in our previous tests of independence. Therefore, for each cell the "expected" or theoretical frequency is the product of the corresponding column and row totals divided by the grand total. For instance, the expected theoretical frequency of boys in High School A is $(45)(87)/159 = 24.62$.

Proceeding with the general formula for χ^2 either equation 10.1 or 10.3 we find $\chi^2 = 2.33$. Table 10.4 is a 4×2 table, that is, it has four rows and two columns. There are thus $3 \times 1 = 3$ degrees of freedom, according to the rule mentioned previously. For 3 d.f. χ^2 would have to be 7.85 before we would reject the hypothesis at the 5 percent level. It is reasonable to hold, therefore, that the four samples of subjects are from a population which is *homogeneous* as to the distribution between the sexes.

10.7 THE $R \times 2$ TABLE

The last example involved a dichotomy, boy *versus* girl, on one criterion of classification. To test homogeneity it is not necessary that the different sets of observations be classified into only two categories. For instance, the samples of students from the four high schools could have been distributed according to how they responded to some item on a questionnaire. The response categories could be: (*a*) favor, (*b*) oppose, (*c*) undecided, or we might have had many more than just the four high schools in our table. However, having only two columns, Table 10.4 may be used to illustrate some special computation techniques convenient for $R \times 2$ tables.

For example, χ^2 can be computed from

$$\chi^2 = \frac{1}{nn'} \sum \frac{1}{a + a'} (an' - a'n)^2 \tag{10.5}$$

where a and a' represent the pair of observed frequencies in any row, and n and n' the corresponding column totals. For instance, in Table 10.4, $n = 87$; $n' = 72$; and in the first row, for High School A, $a = 22$; and $a' = 23$. Substituting the appropriate values for each row and summing over all four rows, we obtain

$$\chi^2 = \left(\frac{1}{6,264}\right)(14,589.8) = 2.33$$

An alternative method makes it possible to compute χ^2 directly from proportions. For instance, we may be interested in knowing the proportion of boys in each of the high schools. These are given in the fourth column of Table 10.5 We see that the proportions of boys varies from slightly more than 48.5 percent in High School C to over 61 percent in High School D. In fact, our test of homogeneity was a test of the hypothesis that these proportions were samples from a single population.

In our previous notation we define $p = a/(a + a')$, and $\bar{p} = n/(n + n')$, and

$$\chi^2 = \frac{1}{\bar{p}\bar{q}} (\Sigma ap - n\bar{p}) \tag{10.6}$$

By multiplying the figures in columns 3 and 4, and summing, we find that $\Sigma ap = 48.1810$, as shown in column 5. We find $\bar{p} = 87/159 = .547170$. As in our notation for the binomial, we define $\bar{q} = 1 - \bar{p} = .452830$. Substituting in formula 10.6, we have

$$\chi^2 = \frac{48.1809 - 47.6038}{.24777} = 2.33$$

TABLE 10.5

SAMPLES OF STUDENTS IN FOUR HIGH SCHOOLS AS PROPORTION OF BOYS

High School	Number of Students Total	Number of Students (a) Boys	(p) Proportion Boys	(ap)
(1)	(2)	(3)	(4)	(5)
A	45	22	.488889	10.7556
B	35	21	.600000	12.6000
C	35	17	.485714	8.2571
D	44	27	.613636	16.5682
Total	159	87	.547170	48.1809

10.8 CHI SQUARE AND THE BINOMIAL DISTRIBUTION

In this chapter we have used notation and terminology which we used in Chapter 5 in discussing the binomial distribution. This is particularly apt in the applications of χ^2 to problems involving one degree of freedom. We recall that in Section 5.4 the binomial distribution was introduced in connection with a hypothetical teacher-prediction problem in which we tested the hypothesis that the observed number of successes, $X = 8$, of the teachers' predictions was from a binomial for which $p = 1/2$ and $n = 10$. In Section 6.5 we saw that the binomial was fairly closely approximated by the normal distribution. The actual binomial probability of eight successes or better was found to be .0547, and the normal probability .0571.

The problem may be reconstructed in the following table:

	f_0	f_t
Successes	$X = 8$	$np = 5$
Failures	$(n - X) = 2$	$nq = 5$
Total	$n = 10$	$n(p + q) = 10$

As was noted in Section 6.5, values of 8 *or more* would be better measured in a *continuous* distribution as values exceeding 7.5, and values of 2 or less as values less than 2.5. Similarly, we should make a *correction for continuity* when we use the continuous χ^2 function, just as we do when we use the normal distribution to approximate binomial probabilities. The continuity correction is rarely needed in χ^2 computations for more than 1 degree of freedom. However, it is best practice always to apply the correction in computing χ^2 for 1 degree of freedom.

Our two-by-two table is thus revised as follows:

	f_0	f_t
Successes	$X = 7.5$	$np = 5$
Failures	$(n - X) = 2.5$	$nq = 5$
Total	10	10

A direct application of equation 10.1 yields

$$\chi^2 = \frac{(7.5 - 5)^2}{5} + \frac{(2.5 - 5)^2}{5} = 2.5$$

Reference to Appendix H shows that a χ^2 of 2.5 would have a probability slightly greater than .10, for the table shows that $\chi^2_{.90} = 2.706$. Actually, $P(\chi^2 > 2.5) = .114$.

The χ^2 test is a *two-tailed test*, and the probabilities we get from the tabled values are probabilities for getting a deviation from the theoretical value as great or greater in *either direction*. The hypothesis tested in Chapters 5 and 6 was a *one-tailed* hypothesis. We were asking the question: What is the probability, by chance alone, of observing eight or more successes from the hypothetical binomial?

Therefore we should halve the tabular probability because we are interested in only one tail. One-half of .114 is .057. This is the same as the approximation of Section 6.5, using the normal distribution and quite close to the exact binomial probability of .0547.

The exact relationship of χ^2 to the normal deviate should be noted for the special case of one degree of freedom. Referring to the above two-by-two table, and using symbols instead of numbers, we compute χ^2 as follows:

$$\begin{aligned}\chi^2 &= \frac{(X - np)^2}{np} + \frac{(n - X - nq^2)}{nq} \\ &= \frac{q(X - np)^2}{npq} + \frac{p(n - X - n + np)^2}{npq} \qquad (10.7) \\ &= \frac{(p + q)(X - np)^2}{npq} = \frac{(X - np)^2}{npq}\end{aligned}$$

This value of χ^2 is *exactly* the *square* of the normal deviate, $z = \dfrac{X - np}{\sqrt{npq}}$, which was computed in Section 6.5 in the normal approximation to the binomial.

Remembering that the χ^2 test is a two-tailed test, tabled values tell us the probability of a chance discrepancy *in either direction.* Taking this into account, we should be able to construct a χ^2 distribution for one degree of freedom from the table of the cumulative normal function. By squaring *two-tailed* critical values of z we should get critical values of χ^2 for one degree of freedom. For instance, for a 20 percent two-tailed test we would use $z_{.90} = 1.28$. Its square is 1.64, the value of $\chi^2_{.90}$. Similarly, the 5 percent critical value for the normal distribution (two-tailed) is $z_{.975}$. In Appendix D we find that this is 1.96. Squared, it is 3.84, the value of $\chi^2_{.95}$ in Appendix H.

We may examine in Table 10.6 one other application of χ^2 which is related to the binomial. It is similar to the problem of Table 10.5, but is especially convenient here for demonstrating another feature of χ^2 because it has samples of equal size.

An experimenter administered a performance test to 6 sets of $n = 25$ subjects, each set from a different training establishment. He was interested in knowing whether he could reasonably consider the 6 samples as coming from a homogeneous population. His hypothesis was that the probability of success on the performance test was the same in the 6 populations supplying the samples. This is a binomial hypothesis, the experimenter wishing to know whether the observed frequencies, X_i, shown in column 3, might reasonably have come from the same population. In column 4 we see that the actual "proportions succeeding" vary from .28 to .76 among the 6 samples and the average proportion of success over all samples is .52667.

For these data χ^2 may be computed in the usual way, using equations 10.1 or 10.3 and the values X_i and $n - X_i$ in a two-by-six table. Equation 10.5 could also be used. Either approach will show that χ^2 is 20.35. We

TABLE 10.6

RESULTS OF PERFORMANCE TEST IN SIX EXPERIMENTS, 25 SUBJECTS EACH

Experiment Number	Number of Subjects, n	Actual Number Succeeding, X_i	Proportion Succeeding, p_i	Theoretical Number Succeeding, $n\bar{p}$	Chi Square χ^2
(1)	(2)	(3)	(4)	(5)	(6)
1	25	8	.32	13.167	4.28
2	25	19	.76	13.167	5.46
3	25	12	.48	13.167	.22
4	25	7	.28	13.167	6.10
5	25	18	.72	13.167	3.75
6	25	15	.60	13.167	.54
All experiments	$nk = 150$	$\Sigma X_i = 79$	$\bar{p} = .52667$	—	20.35

will compute χ^2 by another method which utilizes the particular arrangement of information in Table 10.6 designed to display the binomial aspects of the problem.

First we observe that the theoretical frequency is derived from the table as a whole as $n\bar{p} = 13.167$, the expected number succeeding in a sample of 25 when the binomial probability of success is $\bar{p} = .52667$. We note further that $\bar{p} = \Sigma X_i/nk$ and that $n\bar{p} = \Sigma X_i/k = \bar{X}_i$, where k is the number of separate samples. The expected frequency, $n\bar{p}$, in column 5 is thus the mean of the observed frequencies, X_i, in column 3, but it may be viewed as the mean of a binomial with $n = 25$ and $p = .52667$. The variance of this binomial is $\sigma^2 = n\bar{p}\bar{q} = (25)(.52667)(.47333) = 6.232$. The square root of this is $\sigma = 2.496$, the standard deviation of the binomial distribution.

If the hypothesis is indeed true and the values of X_i in column 3 represent sample deviations from the expected value, we could divide the deviates by 2.496 to express them in standard score form. The results of squaring and summing these deviates (in standard score form) is shown in column 6. The sum is seen to be precisely the χ^2 which was found by other methods of computation. Algebraic proof can be used to show the equivalence. The following expression of this method of arriving at χ^2 is informative:

$$\chi^2 = \sum_{i=1}^{k} \frac{(X_i - n\bar{p})^2}{n\bar{p}\bar{q}} \qquad (10.8)$$

$$= \frac{\Sigma(X_i - \bar{X}_i)^2}{\sigma^2} = \frac{\Sigma x_i^2}{\sigma^2}$$

By examining equation 10.8 and reviewing the numerical procedure of the example, we see that the greater the value of the sum of squares of the deviates, Σx_i^2, the greater the value of χ^2. In the example, the sum of squares of deviations is great enough to produce a χ^2 of 20.35 which exceeds the 1 percent point, thus casting doubt upon the hypothesis.

Equation 10.8 parallels a definition of "continuous" χ^2 which is of considerable practical and theoretical significance. In fact, equation 10.2 is the distribution of the sum of squares of ν independent random values of z, the normal deviate with zero mean and unit variance. We may write

$$\chi^2 = \frac{\Sigma x^2}{\sigma^2} = \Sigma z^2 \tag{10.9}$$

where the number of degrees of freedom is the number of independent deviates. The similarity of equation 10.8 and 10.9 should be noted.

10.9 THE DISTRIBUTION OF THE VARIANCE

Since equation 10.9 involves the distribution of the sum of squares of deviations, we should be able to use it for the sampling distribution of a sample variance. A slight modification of equation 10.9 is

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}; \quad \text{d.f.} = n - 1 \tag{10.10}$$

Hence, a sample variance, s^2, is distributed as $\chi^2 \left(\frac{\sigma^2}{n-1}\right)$.

Equation 10.10 permits testing hypotheses concerning variances and establishing confidence intervals for variances. Suppose that a sample of 25 is found to have a variance of 12, and we wish to test the hypothesis that this is a random sample from a universe whose variance is 9, ($H : \sigma^2 = 9$). Substituting in equation 10.10, we compute $\chi^2 = (24)(12)/(9) = 32$. From Appendix H we find that for 24 d.f., $\chi^2_{.80} = 29.55$ and $\chi^2_{.90} = 33.20$. Hence the probability of getting a value of χ^2 as large or larger than 32 lies between .10 and .20. We therefore consider it reasonably likely that a variance of as much as 12 from a sample of $n = 25$ would occur by random sampling from a population whose variance is 9.

We can use the tabled values of the χ^2 distribution, as in Appendix H, to establish confidence intervals for the variance, σ^2. For instance, in a sample of 20, the variance was computed to be 34.51. We may compute 96 percent confidence limits from $(n-1)s^2/\chi^2_{.02}$ and $(n-1)s^2/\chi^2_{.98}$. For the 96 percent confidence interval we are choosing the value of χ^2 at each end of the distribution which marks off a 2 percent tail. With 19 d.f., the limits for χ^2 are 33.687 and 8.567, from Appendix H. We find the limits of σ^2 to be (19)(34.51)/(33.687) and (19)(34.51)/(8.567) or 19.46 and 76.54.

10.10 THE FOURFOLD TABLE

There are several aspects of χ^2 applications which are peculiar to the two-by-two table. The following tabulation gives frequencies of the 35 students in High School B (from the material in Appendix G) according to two criteria of classification, whether the student intends to go to college, and sex.

	College	Noncollege	
Male	(*a*) 13	(*b*) 8	21
Female	(*c*) 5	(*d*) 9	14
	18	17	35

We can calculate χ^2 directly according to equation 10.1, computing theoretical frequencies from the marginal totals in the table. For example, in cell (*a*) the theoretical frequency is $(18)(21)/(35) = 10.8$. As has been suggested previously, χ^2 is a better approximation if we correct for continuity in cases involving only *one degree of freedom*. We can correct for continuity either by reducing the differences in our computation each by one-half, or by increasing the smallest observed frequency by one-half and adding or subtracting one-half to the other observed frequencies, so that the marginal totals are unchanged.

The correction for continuity is particularly important when any expected frequency is small, but should always be used with a single degree of freedom.

In the foregoing table the theoretical frequencies for cells *a*, *b*, *c*, and *d* are, respectively, 10.8, 10.2, 7.2, and 6.8. The differences between the observed and theoretical frequencies *without correction* are the same, 2.2 (except for sign). *With correction*, they are 1.7. Squaring the differences *after correction*, dividing by the theoretical frequencies, and summing, we find χ^2 to be 1.38. For one degree of freedom this is not large enough to cause us to reject the hypothesis, the probability being between .20 and .30. Although the table shows a greater proportion of boys who intend to go to college, we accept as tenable the hypothesis that this is merely a random variation from a universe in which college-going intentions of the two sexes are the same. As is seen in Chapter 11, this is identical to a test of significance of the difference between two percentages or two proportions.

Computation for fourfold tables may be shortened by means of the formula

$$\chi^2 = \frac{\left(|ad - bc| - \frac{n}{2}\right)^2 n}{(a + b)(a + c)(b + d)(c + d)} \tag{10.11}$$

The numerator involves the absolute value of the difference between the products of diagonally opposite cells. The term, $n/2$, is the *correction for continuity*, and always *reduces* the numerator. In the present example the difference between the cross-product terms is 77, and 17.5 is subtracted for the continuity correction. The remainder ot the application of equation 10.11 is left to the reader. The fraction should come to 123,908.75/89,964 = 1.38.

A series of chi squares computed from independent experiments may be added together, and their sum will be a χ^2 with a number of degrees of freedom equal to the sum of the degrees of freedom of the separate experiments. The resulting χ^2 may be used to test the significance of the outcome of the set of experiments taken together. If each of the several experiments to be combined consists of one degree of freedom, it is best *not* to use the correction for continuity. It is an overcorrection and can accumulate to a seriously lowered χ^2 if a number of chi squares, each with one degree of freedom, are added to form a total χ^2.

Suppose that the following is the table of frequencies resulting from one of six experiments:

	Response *A*	Response *B*
Experimental	14	4
Control	12	10

By the usual methods (without correcting for continuity) χ^2 may be found to be 2.35. Since the probability of this value of χ^2 is greater than .10, it is not sufficient evidence for rejecting the hypothesis of independence of response and experimental treatment at either the .01 or .05 levels.

Similarly chi squares are computed for the other five experiments and are found to be 2.51, 3.01, 4.02, 1.75, and 3.58. The sum of the six chi squares is 17.22. Since this exceeds the one percent critical value for six degrees of freedom, the hypothesis of no difference in response of experimental and control treatments would be rejected on the strength of the combined experiments at the .01 level.

10.11 CAUTIONS IN THE USE OF CHI SQUARE

There has been considerable exchange among scientific workers concerning the use of χ^2 in practical work. In the fields of psychology and education it is used much more extensively than it was 25 years ago. There has thus been opportunity for observing the various misuses which can be made of it. Misuse is usually attributable to a lack of understanding of basic principles concerning χ^2. It must be remembered that the "frequency formulas" are merely approximated by the continuous function. χ^2 itself, that is, the function for which values are tabled, is defined as the sum of squares of normal deviates expressed in standard form.

Among the cautions which must be exercised in using χ^2 is making sure that the "events" or "measures" to be analyzed are derived from random samples. It is advisable to use the correction for continuity in cases involving only one degree of freedom. It may sometimes be considered necessary to determine exact probabilities abandoning χ^2 calculations (see reference 5). When the χ^2 has degrees of freedom greater than one, we apply no correction for continuity. For very small theoretical frequencies χ^2 is not generally a good approximation. For further information which will help avoid improper applications of χ^2, consult the Cochran reference (1) at the end of this chapter.

There is a tendency for an inexperienced person to apply χ^2 to data expressed as percentages, proportions, or scales of values and not frequencies. As explained in previous sections, χ^2 is approximate for dealing with proportions (or percentages). However, it is essential that the proper formulas be used.

Finally, among possible causes of misuse of χ^2, there is the danger of computational error. Care to achieve computational accuracy is, of course, a caution which should be exercised in all statistical work. An improper entry in a χ^2 table, the wrong computation of degrees of freedom, a difference not squared, or figures improperly divided, all of these things happen frequently unless all operations are carefully checked and verified.

EXERCISES

1. One of the ways of testing the randomness of a table of random numbers is to see if the digits 0 to 9, inclusive, appear in acceptably equal frequencies. Select a sample

of random numbers from the table in Appendix D and test the hypothesis that each digit occurs with equal frequency in the table. What is the effect of size of sample upon the sensitivity of the test? Can you think of other matters which should be tested if we are to verify the "randomness" of a table of numbers?

2. From the information of Ex. 3 of Chapter 3 record Q_1, Md, Q_3 for the CTMM scores in Appendix A. These mark the limits of four classes of score for which the probability is the same. Using the table of random numbers in Appendix B, draw samples of size 20 (*with replacement*) from the CTMM scores of Appendix A. For each sample tabulate the observed distribution in the four quartiles against the expected frequencies of 5 each. Compute χ^2 for each sample. Make a frequency distribution of observed values of χ^2 obtained from all members of the class. Compute the ninety-fifth percentile and the ninety-ninth percentile of this distribution. Compare this distribution with the distribution in Fig. 10.1. Compare critical values with those in the table of Appendix H.

3. When is it advisable to correct for continuity when using χ^2 in testing a hypothesis? Why is this correction necessary?

4. Assume that the 159 observations in Appendix G are random observations. At the .05 level, test the hypothesis that there is no difference between the proportion of boys going to college and the proportion of girls going to college.

5. From the data in Appendix G for High School B, would it be reasonable to assume that the number of senior boys and the number of senior girls in High School B are the same?

6. In a study comparing reading habits of retarded and normal school children, several measurements were made of physical characteristics thought to have a bearing on learning to read. One of these was eyedness. Eyedness classifications for the 20 retarded and the 30 normal children were as follows:

	Eyedness	
Group	Left	Right
Retarded	13	5
Normal	25	5

Using χ^2, test the hypothesis of no difference in eyedness between the two groups.

7. Of 100 digits drawn from a table of numbers, 60 were found to be even numbers. At the .05 level, test the hypothesis that the table contains equal numbers of odd and even digits.

8. From results of Ex. 8 in Chapter 6 test the hypothesis that the California test scores are normally distributed.

9. Using the methods of this chapter, test the hypothesis of Ex. 6, Chapter. 7.

10. Of 557 boys in a statewide sample of high-school seniors, 113 listed mathematics as their favorite subject. Of the 643 girls in the sample, 45 listed mathematics. Is there a relationship between sex and preference for mathematics among high-school seniors in the population sampled? (Use a .01 level of significance.)

11. In what ways is the χ^2 distribution related to the normal distribution?

12. From the data given in the table decide whether or not there is a relationship between achievement in high school and occupation of father.

Achievement Average	Occupation of Father						
	A	B	C	D	E	F	G
High	36	89	22	81	17	31	97
Average	19	59	18	84	22	28	99
Low	7	30	6	44	15	29	89

13. In a survey of 1,586 city-school superintendents one item of information was "position held immediately preceding first superintendency." The following is a tabulation of the 1,586 returns by three city-size classes and by four categories of position held immediately preceding first superintendency.

Immediately Preceding Position	City Size			Total
	2,500–2,999	10,000–29,999	30,000–and over	
Teacher	94	59	19	172
High-school principal	408	291	129	828
Other administration	191	185	118	494
Other	41	28	23	92
Total	734	563	289	1586

Are size of city in which employed and type of position immediately preceding a superintendent's first superintendency independent? Specify the level chosen and the hypothesis to be tested.

14. The variance of a sample of size 15 was found to be 34.8. What are the 98 percent confidence limits for σ^2?

15. The standard deviation of a normal infinite population is known to be 8. Suppose a very large number of samples of size 17 is to be drawn randomly from this population and that the variance is computed for each sample. What is the expected value of the mean of the distribution of these sample variances? What is your best estimate of what the median would be of this distribution? How do you account for any differences between the expected mean and the expected median? Would the distribution be normal? Would the shape of the distribution be different for a different sample size? What is your estimate of the range of the middle 90 percent of sample values of the variance? Of the standard deviation? What other percentile values of the distribution of the sample variances can you estimate easily from tables in this book?

16. A follow-up study was conducted by questionnaires sent to all the graduates in three annual graduating classes of a group of high schools. The following *percentages* were reported to show the difference in the marital status between male and female graduates:

Marital Status	Male	Female	Total
Married	13.2	33.1	24.0
Single	83.3	64.0	72.8
No reply	3.5	2.9	3.2

(*a*) Why is it not possible with no more than the above information to use χ^2 for a test of homogeneity?

(*b*) Suppose that in the appendix of the report it is found that the foregoing is based upon a total of 250 questionnaire returns. How would you reconstruct the table of frequencies from which the above percentages were derived?

Hint: Let M = number of males and F = number of females. Then

$$.132M + .331F = (.240)(250)$$

and

$$.833M + .640F = (.240)(250)$$

Solve the above simultaneous equations for M and F.

17. Suppose that in the study of the previous exercise there were 15 married males, 45 married females, 95 single males, 87 single females, and 4 of each sex from whom there was no reply.

(*a*) How would you use χ^2 in the analysis of the data? Define your population, state your hypothesis, specify the level, and carry out the application of χ^2 which you think is appropriate.

(*b*) In a problem such as this how do the " no reply" frequencies disturb your analysis and interpretation?

(*c*) What statistics other than χ^2 might be useful in the analysis of these data?

(*d*) Suppose that you are told that the 250 graduates were the only ones responding to the questionnaire from a random sample of 1,000 to whom it had been mailed. What effect would this have upon your interpretation of the data and the use of χ^2?

18. Fifty subjects are randomly assigned to 5 different methods of training to operate a complicated piece of equipment in military aircraft. A measure of efficiency in training is based upon length of time required to reach a prescribed standard of proficiency in a flight test. Subjects in the four groups were classified on this measure into two classes, those above and those below the median of the 50 measures as follows:

	A	B	C	D	E
Above *Md*	6	2	7	6	4
Below *Md*	3	9	4	4	5

At the .05 level, test the hypothesis that the medians of the 5 methods groups are the same and equal to the general median.

19. Subjects in an experiment are classified in two treatment groups and into "success" and "fail" categories on the outcome of a criterion test. The result is a two-by-two

table of frequencies. χ^2 is used to test the hypothesis that there is no difference between the two groups in the proportion of success. This experiment is repeated independently six times. The chi squares are found to be 4.23, 1.25, .43, 1.73, 3.25, and 2.41. Combining the results of the 6 experiments, test the hypothesis at the .05 level that there is no difference in the two treatments.

REFERENCES

1. Cochran, William G., "The χ^2 Test of Goodness of Fit," *Annals of Mathematical Statistics*, 23: 315-45, September 1952.

2. Edwards, Allen L., *Statistical Methods for the Behavioral Sciences*, New York, Rinehart and Co., 1954, Chapter 18.

3. Freund, John E., *Modern Elementary Statistics*, New York, Prentice-Hall, 1952, Chapters 15 ard 16.

4. Mood, Alexander M., *Introduction to the Theory of Statistics*, New York, McGraw-Hill Book Co., 1950, pp. 273-282.

5. Walker, Helen M., and Joseph Lev, *Statistical Inference*, New York, Henry Holt and Co., 1953, Chapter 4.

6. Wert, James E., Charles O. Neidt, and J. Stanley Ahmann, *Statistical Methods in Educational and Psychological Research*, New York, Appleton-Century-Crofts, 1954, Chapter 9.

CHAPTER 11

Comparing Two Samples

Applications of sampling distribution theory to the testing of hypotheses on the basis of a single sample have been considered in previous chapters. Very frequently in educational work questions are asked which involve comparing two samples. We have already found chi square to be useful in such problems when data are expressed in frequencies, but this is not often the case. For example, we may wish to compare a sample of "good readers" and a sample of "poor readers" on various measures concerning the reading process. In an educational experiment there may be an "experimental" group of students who have received instruction under some novel or experimental method. We wish to compare the experimental group with a "control" group, that is, another sample of pupils whose training has been offered under conventional methods. A sample of salaries of professors in one type of higher educational institution is found to exceed the mean of a similar sample from another type of institution. Is a difference in the same direction to be expected in the total populations of the two classes of institution?

A sample of pupils, from a given population, may be tested at two different times, such as before and after an experiment. The investigator here would be interested in the difference in performance resulting from the experiment. This test differs from the previous ones in that the two sets of measures are *not independent*. The same is true of scores from two different sets of individuals if they are paired or "matched" in some manner before experimentation. The approach to the comparison of "independent" groups is the same as that of "paired" groups. An early step in planning statistical tests of differences between two samples, for reasons which will be explained in this chapter, is the decision as to whether pairing is appropriate.

We shall see that our earlier discussions of statistical significance and statistical inference apply to such problems. However, we first digress to consider variances of sums and differences.

11.1 THE VARIANCE OF SUMS AND DIFFERENCES

We first consider the variance of a difference between two sets of paired or "correlated" measures. A simple numerical illustration appears in Table 11.1. Here we have five pairs of measures, X_1 and X_2. There are a number of ways in which measures can be paired. For instance, we may have two measures on the *same* individual—for individual "a" 8 and 10, and individual "b" 7 and 13, respectively, and so on. These two measures may be scores on the same test administered before and after an experiment to the five subjects, or it may be two different tests administered at the same sitting. Another possibility, which this situation represents, is the same test administered to ten *different* individuals who were arranged in some manner in five pairs before testing. The two members in each pair were alike in some respect, such as mental age or grade. The X_1 individuals and the X_2 individuals are, however, classed distinctively in their respective groups. For instance, the X_1 individuals may be those who are high-school graduates and the X_2's those who are not, or the first group of measures may be from subjects in a control group and the second set of measures may be those derived from five subjects in an experimental group.

Our immediate object is to observe the relationship between the variance of the *differences* and the variances of the two sets of scores. These differences are shown in the last column of Table 11.1. They may represent gains of the same five individuals in an experiment or measures of the superiority of five individuals over their respective mates with which they have been paired. We first observe that the means of the two sets of scores are respectively 6 and 12, and that the mean of the differences is 6, the difference between the two means. If we let $D = X_2 - X_1$, and $\bar{D} = \Sigma D/n$, we may write

$$\bar{D} = \frac{\Sigma(X_2 - X_1)}{n} = \frac{\Sigma X_2}{n} - \frac{\Sigma X_1}{n} = \bar{X}_2 - \bar{X}_1$$

TABLE 11.1
THE DIFFERENCES BETWEEN FIVE PAIRS OF MEASURES

Pair	First Measure, X_1	Second Measure, X_2	Difference, $D = X_2 - X_1$
a	8	13	5
b	7	14	7
c	6	10	4
d	5	12	7
e	4	11	7
Total	30	60	30

Knowing the three means for the three columns of figures in Table 11.1, we find it easy to compute the deviations, the squares of deviations, and the sums of each. In the last column, the deviations of the differences, $d = D - \bar{D}$, are $-1, +1, -2, +1$, and $+1$. The sum of their squares is $\Sigma d^2 = 8$.

We now note that the deviation of a difference is equal to the difference of the respective deviations. For any single pair,

$$\begin{aligned} d = D - \bar{D} &= (X_2 - X_1) - (\bar{X}_2 - \bar{X}_1) \\ &= (X_2 - \bar{X}_2) - (X_1 - \bar{X}_1) \\ &= x_2 - x_1 \end{aligned}$$

Squaring and summing over all pairs, we find the sum of squares of deviations to be

$$\begin{aligned} \Sigma d^2 &= \Sigma(x_2^2 - 2x_2x_1 + x_1^2) \\ &= \Sigma x_1^2 + \Sigma x_2^2 - 2\Sigma x_1x_2 \end{aligned} \tag{11.1}$$

The sums of squares of deviations of the two sets of scores are respectively 10 and 10, and the sum of squares of deviations of differences is 8. It is simple to find the sum of the products of deviations. The product of deviations for pair a is 2, for pair b is 2, and so on. We find $\Sigma x_1x_2 = +6$. Substituting in the above identity, we find that $\Sigma d^2 = 8 = 10 + 10 - 12$.

If we divide equation 11.1 through by $(n - 1)$ we find that the variance of the differences is equal to the sum of the variances of the two sets of measures minus twice their covariance. That is,

$$\begin{aligned} s_D^2 &= s_1^2 + s_2^2 - 2\ \text{cov}_{12} \\ &= s_1^2 + s_2^2 - 2s_1s_2r_{12} \end{aligned} \tag{11.2}$$

Thus the variance of the difference between pairs depends upon their correlation, as well as upon their respective variances. It is worth verifying equation 11.2 from the data of Table 11.1 to see how this relationship operates. The variance of the differences is found to be 2.0. The variances of the two sets of measures are each 2.5. Twice the covariance is $2(6/4) = 3.0$.

It is easy to find the observed correlation coefficient, .60, a fair relationship between the two variables. We see that the *higher* the correlation the *lower* the variance of the differences. Had there been perfect correlation, what would the variance of the differences have been? An empirical way to work this through would be to rearrange the second set of measures so that the correlation would be perfect and see what happens to the differences. Little computation is required to see that if the X_2 measures had been for pairs a, b, c in order 14, 13, 12, etc., that (*a*) the correlation would have been perfect and (*b*) all the differences would have been the same. Each D would have been precisely 6. Since all the differences would be the same in this particular case, the *variance* of these differences would, of course, be zero. From the data of the table thus revised we could substitute in equation 11.2. The covariance, that is, the value for the $s_1s_2r_{12}$ term, would be 2.5. Twice this would be 5.0, and from equation 11.2 we would find $s_D^2 = 2.5 + 2.5 - 5.0 = 0$.

If we now imagine a situation in which there is exactly zero correlation between the two sets, the third term of equation 11.2 will drop out and

$$s_D^2 = s_1^2 + s_2^2 \tag{11.3}$$

In this case the variance of the differences is precisely the sum of the variance of the two sets of measures. Where correlation is between 0 and + 1.0 the variance of the differences will be some place in between zero and the sum of the variances, depending upon the degree of correlation. The higher the correlation, the smaller the variance of the differences.

Another possible situation to contemplate is that in which the correlation is negative. By actually manipulating the figures of Table 11.1, or imagining such manipulation, such that the highest score in one set is paired with the lowest score in the other, and so on, so that there is perfect *negative* correlation, we can visualize what effect this would have on the differences. It is readily seen that they would be as different as possible, ranging all the way from +2 to +10. A negative correlation would cause the last term of equation 11.2 to be positive. A *negative* correlation makes the variance of differences larger.

A similar development may be made for the variance of *sums* of paired scores, $X_1 + X_2$. An occasion for interest in this would be finding the variance of the total score on a test from scores on two parts of a test given to the same n subjects. We assume here that the sums of the scores on the two parts of the test equal the score on the whole test. An approach similar to that above will readily show that the variance of the sum is

$$s_{x_1+x_2}^2 = s_1^2 + s_2^2 + 2s_1s_2r_{12} \tag{11.4}$$

As in equation 11.2, this shows us that the variance of combined scores is dependent upon the respective variances *and the correlation* between the two sets. However, in this instance the third term is positive.

An important observation to make is that the variance of sums and the variance of differences both depend upon the sums of the two variances, s_1^2 and s_2^2. Unless we have acquainted ourselves with these relationships, we might suppose that the variance of differences would be related to the difference between the variances. This, we have seen, is not true. A correlation term is involved in both equations 11.2 and 11.4. In one case positive correlation *reduces* the variance of differences; in the other positive correlation *increases* the variance of sums.

If the correlation is zero and only then,

$$s^2_{x_1+x_2} = s_1^2 + s_2^2 \tag{11.5}$$

We note from equations 11.3 and 11.5 that, if the correlation is zero, the variance of differences and the variances of sums would be identical and equal to the sum of the variances of the two sets.

11.2 THE SIGNIFICANCE OF A DIFFERENCE BETWEEN TWO MEANS—PAIRED MEASURES

The most direct method of testing the significance of the difference between means of paired measures (matched or correlated measures as they are sometimes called) requires only the application of the principles of Chapter 7 concerning the mean, as the following example shows.

TABLE 11.2

DIFFERENCES BETWEEN PAIRS OF SCORES

Pair	Experimental Group, X_1	Control Group, X_2	Difference $D = (X_1 - X_2)$
(1)	(2)	(3)	(4)
a	18.4	14.7	3.7
b	13.9	10.5	3.4
c	9.5	10.4	−.9
d	15.2	13.8	1.4
e	16.3	14.6	1.7
f	14.0	9.2	4.8
g	18.2	16.7	1.5
h	15.8	12.5	3.3
i	13.2	6.3	6.9
j	16.1	13.1	3.0
Total	150.6	121.8	28.8
Mean	15.06	12.18	2.88
Squares	2,329.08	1,569.18	123.70

In Table 11.2 are pairs of scores on an arithmetic test for 10 subjects, "matched" on intelligence test scores, previous achievement in arithmetic, and a pretest. There were thus 20 subjects, 10 in an experimental group which received special remedial instruction in arithmetic, and 10 in a control group which received no special instruction. Our observed means are 15.06 for the experimental group and 12.18 for the control group. We wish to test the null hypothesis, $H : \mu_1 = \mu_2$, that the difference between the population means is *zero*. We will do this by deriving from the two sets of measures, X_1 and X_2, a third set of measures $D = X_1 - X_2$. We now visualize our problem as simply that of a hypothesis concerning the universe of values from which were drawn the 10 *differences* shown in column 4 of Table 11.2. True, the figures with which we are dealing are actually differences, but we treat them as we would any set of measures, giving no further attention to the data in columns 2 and 3.

The procedure is simply that of Section 7.10. The hypothesis is that the universe of measures (differences) has a mean of zero. Our sample of differences is one of a large number of such samples of size 10 which might be obtained by performing a large number of experiments similar to the one we are considering. We do not know the variance of the sampling distribution of sample means, such as the observed sample mean, $\bar{D} = 2.88$, nor do we know the variance of the universe of differences, but we may apply equation 7.4 directly and estimate the variance of the mean difference as

$$s_{\bar{D}}^2 = s_D^2/n = (\Sigma d^2)/n(n-1) \tag{11.6}$$

This turns out to be

$$s_{\bar{D}}^2 = (123.70 - 82.94)/(10)(9) = .4529$$

The square root is the estimated standard error of the mean difference. It is $s_{\bar{D}} = .673$. The appropriate statistic, since we are *estimating* the standard error, is

$$t = (\bar{D} - 0)/s_{\bar{D}} \tag{11.7}$$

In order to test the hypothesis of no difference in the universe, we compute $t = 2.88/.673 = 4.28$. For 9 d.f. the probability of t is less than .01. We would, therefore, reject the hypothesis at the 1 percent level.

The chief feature of the foregoing procedure is that we first derive the differences of the pairs of scores. Then we treat the differences as a sample set of measures and proceed as in Chapter 7 in testing a hypothesis about the population mean.

It would not be necessary to limit ourselves to testing the hypothesis that the difference between the two universe means is zero (or that the universe mean difference is zero). We might wish to test other hypotheses. For instance, if measures were in *age scores or grade scores* the conditions of the experiment might be such that the experimenter would wish to test the hypothesis that the difference is half a year or half a grade in achievement. In this case the numerator of equation 11.7 would be the difference between the *observed* mean difference and whatever value of the population difference was involved in the *hypothesis*. If confidence intervals are desired, we apply equation 7.6.

The above method is commonly used and is reasonably simple to understand. A more complicated method, based on a different but equivalent computation of the standard error of the mean, will be used to demonstrate the importance of the correlation of matched pairs in two groups of measures, X_1 and X_2.

Dividing equation 11.2 by n, we have the following:

$$s_{\bar{D}}^2 = s_{\bar{x}_1}^2 + s_{\bar{x}_2}^2 - 2s_{\bar{x}_1}s_{\bar{x}_2}r_{12} \qquad (11.8)$$

From columns 2 and 3 of Table 11.2, we may compute sums of squares of deviations and sums of products of deviations, and from these the variances, the standard errors, and the correlation coefficient. The results we substitute in equation 11.8 as follows:

$$\begin{aligned} s_{\bar{D}}^2 &= .678 + .952 - 2(.824)(.976)(.732) \\ &= .678 + .952 - 1.177 \end{aligned}$$

As we know, the larger the value of t the less likely it is to occur. In examining equation 11.7 we notice that t varies *inversely* as the standard error of the mean difference (the denominator of t). That is, if we *reduce* the standard error of the difference, we *increase t*. This increases the probability of rejecting the hypothesis of no difference, that is, the probability of finding a significant difference. We now note in our substitution in equation 11.8 that the third term is negative. This term, -1.177, represents the correlation component of the error variance of the mean difference. Without this last term, $s_{\bar{D}}^2$ would be $.678 + .952 = 1.630$. The correlation or covariance term reduces this by 1.177 to .453, which we computed directly from the individual differences. This explains the advantage in experimental design resulting from the matching of subjects. The reason for matching is to reduce the error variance, that is, to increase the precision of the experiment. If the bases upon which subjects are matched is such that pairs will be highly correlated in the measures used in the outcome of the experiment, the "experimental error" is greatly reduced. This increases the chance of finding a significant difference between the two populations if the true difference is not zero. As we shall see later, gain in efficiency by matching because of correlation of pairs must offset an advantage in degrees of freedom which the unmatched design provides.

As pointed out in Chapter 7, statistical hypotheses concerning means may be tested by use of the normal distribution when the *population variance is known* (or when sample size is so large that the normal is an adequate approximation to the t distribution). If the variance of the population of differences is *known*, we may compute

$$z = \bar{D}/\sigma_{\bar{D}} \tag{11.9}$$

If z is larger than 1.96, we would reject the hypothesis of no difference at the 5 percent level. If it exceeds 2.58, we would reject at the 1 percent level. In case our hypothesis is one-sided, we would reject at the 1 percent level if z exceeds 2.33, and so on.

If the variance of the universe of differences is not given, but the two groups are normal and their variances and intercorrelation are known, we may compute the variance of the difference as

$$\sigma^2_{\bar{D}} = \sigma^2_{\bar{x}_1} + \sigma^2_{\bar{x}_2} - 2\sigma_{\bar{x}_1}\sigma_{\bar{x}_2}\,\rho_{12} \tag{11.10}$$

11.3 THE SIGNIFICANCE OF THE DIFFERENCE BETWEEN TWO MEANS—INDEPENDENT SAMPLES

We turn now to the comparison of means of "independent groups." The manner in which subjects are selected in one sample has no effect on the manner in which they are selected in the other. In Appendix G, for instance, are several independent samples of scores on the California Test of Mental Maturity. These are random samples of students in four high schools. Random sampling means that the subject selected, for instance, in high school B, would bear no relation to the subjects in high school C.[1]

Our interest usually is in knowing whether there is a real difference in the criterion measure between an experimental group and a control group, or, in the specific example, between measures in high School B and similar measures from high school C. Hence we test the hypothesis of no difference between the two samples. Our approach is therefore similar to

[1] This does not mean that we might not sample from high school B and high school C in such a manner as to pair them. We might choose randomly one student from a subclass in each school. The subclasses would be determined by some characteristic such as grade average, IQ, chronological age, or social class. There would be two students in pair a from the two different schools who would be in the first subclass, similarly two in subclass b, etc. These would then be "matched" samples, *not* independent random samples now under consideration.

that in the preceding section, although there are differences in the statistical analysis and in the statistical thinking involved.

Let us look first for the standard error of differences between means from independent samples. This is the standard deviation of the sampling distribution of differences between means of a great many pairs of sample means. Referring to equation 11.10 we find that the variance of the difference between means of *independent* groups is

$$\sigma_{\bar{D}}^2 = \sigma_{\bar{x}_1}^2 + \sigma_{\bar{x}_2}^2$$

and the standard error is

$$\sigma_{\bar{D}} = \sqrt{\sigma_{\bar{x}_1}^2 + \sigma_{\bar{x}_2}^2} \qquad (11.11)$$

The third term in equation 11.10 is zero because the groups are independent.

Therefore, if σ_1^2 and σ_2^2 are *known* (or assumed known), and if the distributions are normal, the normal model is exact for testing hypotheses concerning $\mu_1 - \mu_2$, using the ratio $z = (\bar{X}_1 - \bar{X}_2)/\sigma_{\bar{D}}$.

If the two samples are random samples from the *same population* with *known variance*, we may rewrite equation 11.11 as

$$\sigma_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}} = \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \qquad (11.12)$$

Given σ, if the population is normal (or near normal and n is sufficiently large), the distribution of the difference between means is normal (or approximately so) and the normal distribution may be used to determine whether an observed difference between two sample means can be considered tenable under the hypothesis H : $\mu_1 = \mu_2$.

Most frequently the population variance, σ^2, is unknown, and we must estimate it. Such an estimate results from pooling sums of squares of deviations and dividing by the appropriate degrees of freedom. In symbols, this is

$$s^2 = \frac{\Sigma x_1^2 + \Sigma x_2^2}{n_1 + n_2 - 2} \qquad (11.13)$$

We lose a degree of freedom for each of the two sample means. The degrees of freedom are, respectively, $n_1 - 1$ and $n_2 - 1$. Their sum is the denominator in equation 11.13.

The square root of equation 11.13 may now be substituted in 11.12 to yield

$$s_{\bar{x}_1-\bar{x}_2} = \sqrt{\left(\frac{\Sigma x_1^2 + \Sigma x_2^2}{n_1 + n_2 - 2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \qquad (11.14)$$

To make use of this estimate of $\sigma_{\bar{x}_1 - \bar{x}_2}$, we turn again to the t distribution. We saw in Section 7.10 that t is the ratio of a normally distributed variate to the square root of an unbiased estimate of its variance. If the samples are from normal populations with the same variance, $\bar{X}_1 - \bar{X}_2$ is normally distributed about $\mu_1 - \mu_2$. Equation 11.14 is the square root of the unbiased estimate of the variance of the difference between the sample means. Hence,

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_{\bar{x}_1 - \bar{x}_2}} \tag{11.15}$$

is distributed in Student's distribution with $n_1 + n_2 - 2$ degrees of freedom.

The major point of this discussion is that the use of equations 11.14 and 11.15 assumes (*a*) the same variance for the two populations, (*b*) normality of distribution, and (*c*) independence of the two samples. The most common application of this model is testing the hypothesis $H : \mu_1 - \mu_2 = 0$. We will illustrate the procedure with an example.

In Appendix G is a summary of a sample of 13 CTMM scores for high-school boys from high school B who intended to go to college, and a sample of 9 scores on the same test for boys intending to go to college from high school C. We have two "hypothetical" populations. The first is the population of California test scores for all the boys who have gone or will go to high school B and intend to go to college, and a similarly defined population for high school C. We summarize the data from Appendix G as follows:

High School B Sample 1	High School C Sample 2
$n_1 = 13$	$n_2 = 9$
$df_1 = 12$	$df_2 = 8$
$\bar{X}_1 = 69.31$	$\bar{X}_2 = 78.78$
$\Sigma x_1^2 = 1{,}156.77$	$\Sigma x_2^2 = 1{,}703.56$
$s_1^2 = 96.40$	$s_2^2 = 212.94$

These statistics are derived by the routine methods already familiar to us. For instance, $\bar{X}_1 = 901/13 = 69.31$. Also, from the information supplied in Appendix G, we use the formula for computing sums of squares of deviations, $\Sigma x^2 = \Sigma X^2 - (\Sigma X)^2/n$, and we compute $\Sigma x_1^2 = 63{,}603 - (901)^2/13 = 1{,}156.77$.

We note, first of all, that the first sample mean is considerably less than the second sample mean. We note furthermore that the second sample variance is more than twice as great as the first sample variance. This latter point we consider in a later section. The question is: Is there any difference between *the population means*, that is, do boys in high school *B* intending to go to college have the same average CTMM score as boys in high school C intending to go to college? The logic of the method we use is that of the null hypothesis. The question is stated this way: May we consider the two samples, at a stated level of risk, to be from the same population with $\sigma_1 = \sigma_2 = \sigma$ and $\mu_1 = \mu_2 = \mu$. We estimate the standard error of the difference between the means by equation 11.14, pooling the sums of squares of deviations as follows:

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\left(\frac{2{,}860.33}{20}\right)\left(\frac{1}{13} + \frac{1}{9}\right)} = 5.19$$

We substitute in equation 11.15 to find

$$t = \frac{-9.47}{5.19} = -1.82$$

As we have stated the problem, it is a two-sided test. Suppose that the level of risk chosen is $\alpha = .05$. For 20 d.f. the acceptance region for t begins at $-.209$ and ends at $+2.09$. The observed t is, therefore, in the acceptance region. Therefore we do not reject the null hypothesis. In fact, the two samples could reasonably be random samples from the same population. Since the hypothesis is tenable, and since our model includes equality of variances, there is nothing to suggest that we should make a test of the significance of the difference of the two variances. Although our interest in such cases is usually in the *means only*, it is to be emphasized that when a significant result is found with this test, it could sometimes be ascribable to differences in the variances and not in the means. Therefore we may wish to test the significance of difference of variances even when our primary concern is with the means. Also, in experimental work there is sometimes primary interest in testing the significance of differences of variances. We will therefore, discuss the question of comparing variances. Later we will return to the problem of what to do when the assumptions of the model under discussion, that is, *normality* and *homogeneity* of variance, do not hold.

11.4 THE F RATIO AND THE COMPARISON OF VARIANCES

Although we are accustomed to comparing two statistics, such as two means, by computing their differences, it is just as satisfactory in principle to compare two quantities by use of their ratio. This is convenient in testing the equality of two variances. The reason for this is that the distribution of the ratio of two independent sample variances from a normal population is known, and tabled values of the distribution are available. This distribution was called the F distribution or F ratio by Snedecor in honor of R. A. Fisher, who solved the distribution problem. We define

$$F = \frac{s_1^2}{s_2^2} \tag{11.16}$$

The two variances, s_1^2 and s_2^2, are variances of random samples drawn from a normal population. The distribution of F depends upon the number of degrees of freedom in the two variances, $n_1 - 1$ and $n_2 - 1$. Like the t distribution and the chi-square distribution, the F distribution is a whole family of distributions, one for each possible combination of numbers of degrees of freedom.

For each of many such combinations, various percentile points have been tabled. For most purposes, however, only two probability points are required. These are values for $F_{.95}$ and $F_{.99}$. In tables most commonly used they are called the 5 percent and 1 percent points. They are the points below which 95 percent and 99 percent of sample ratios should be expected, respectively. They are thus *one-sided critical values*. That is, they cut off rejection regions respectively, for $\alpha = .05$ and $\alpha = .01$, on the right-hand tail of the distribution. The F ratio is used mostly in analysis of variance techniques. We see in the next chapter that in the typical analysis of variance problem only the one-sided test with the F distribution is used. That is why the table in Appendix I gives only the one-sided critical 5 percent and 1 percent points for various combinations of degrees of freedom.

In order that the proper combination of degrees of freedom may be clearly understood, we will use the scheme of reporting, first, the number of degrees of freedom for the numerator and, second, the number for the denominator. For instance, in an F ratio consisting of a variance in the numerator with 10 d.f. and one in the denominator with 8 d.f., we write: $F(10, 8)$. From the table in Appendix I we may look up the 5 percent point, that is, the value of $F(10, 8)$ which is exceeded 5 percent of the time. It is 3.34. We do this by entering the table first across the top, selecting the proper column for number of degrees of freedom for the numerator, proceeding down this column until we reach the row for the proper number of degrees of freedom in the denominator. In the same table we find that the 5 percent point for $F(8, 10)$ is different. It is 3.07.[1]

[1] More extensive tables for the F distribution are found in several statistics textbooks and in references 3 and 6.

Now suppose that we are interested in testing the hypothesis $H: \sigma_1 = \sigma_2$, and we have available the sample information of the previous section for high school B and high school C scores for the two populations of male students who plan to attend college. The variances of the two samples are, respectively, 96.40 and 212.94. In the F test we use variances, not standard deviations. Thus the hypothesis is really, $H: \sigma_1^2 = \sigma_2^2$, which is equivalent to testing $H: \sigma_1 = \sigma_2$. The appropriate test is two-sided, because a sufficient difference between the two sample variances *in either direction* would lead us to reject the hypothesis.

Now the F distribution, as we have defined it, is the ratio of one random variance to another, and the numerator may sometimes be the smaller. The left-hand tail of the F distribution, therefore, must include fractional values down to zero. To keep the tables compact, only the right-hand tail is tabulated, as appears in Appendix I, where we observe that all critical values of F are greater than 1.00. We obviate the necessity of dealing with the left-hand tail when comparing two variances, by always placing the larger of the two variances in the numerator. In the present case we find $F = 212.94/96.40 = 2.21$. In the table of Appendix I we find the 5 percent point for $F(8, 12)$ to be 2.85 and the 1 percent point to be 4.50. Our observed ratio is less than either of these values.

It is important to remember, however, that the appropriate test is a two-tailed test. Hence the probability of exceeding the 5 percent and 1 percent levels are really 10 percent and 2 percent. In other words, it is necessary to *double the tabular probability* for the two-sided test.

In the two samples from High School B and High School C, though one variance is more than twice the size of the other, we would have no substantial grounds for rejecting the hypothesis that the two samples come from populations of equal variance. It is to be recalled that the t test in the previous section assumes equality of variances in the universes. And as was explained, the t test is an over-all test of equality of variances and equality of means. We would hardly expect, by virtue of the t test (which was not rejected at the 5 percent level), significant differences between the variances. In cases in which significant differences are found by the t test, the F test can be used to examine the possibility that the difference is in variances and not in means.

Though there is rarely a need for it, it is nevertheless possible to determine the lower limits of an acceptance region from tables such as that in Appendix I. As we have seen, the upper critical value, $F_{.95}(8, 12)$, is 2.85. The lower critical value, that is, the point cutting off the lower 5 percent of the distribution, $F_{.05}(8, 12)$, may be computed by finding the reciprocal of the tabled value of F (at the same level) with the degrees of freedom reversed. In the table we find $F_{.95}(12, 8) = 3.28$. Its reciprocal,

$1/3.28 = .305$, is $F_{.05}(8, 12)$. Therefore, in comparing two such variances, we would reject at the 10 percent level if the observed ratio was either less than .305 or greater than 2.85.

We emphasize the assumptions of the F distribution. Its use is strictly appropriate only if the samples are (*a*) *random*, (*b*) *independent*, and (*c*) drawn from *normal* populations. Difficulties in statistical interpretation can sometimes be avoided by taking such matters into account in advance of conducting a statistical study. It is often possible to plan such a study so that it fits into the most convenient and efficient models of statistical inference.

11.5 A NOTE ON EXPERIMENTAL DESIGN

In Section 11.2 we found the means for the two groups in Table 11.2 to be significantly different. We used the t test and we noted that a correlation term reduced the error variance because of matching of the groups. We may glean a notion of the value of matching by comparing the experiment reported in Table 11.2 with a hypothetical experiment with identical results *had there been no matching*. In other words, if we had taken a random sample of subjects and placed them in the experimental group, *group* 1, and an independent sample of ten additional subjects and placed them in *group* 2, we would compare the two groups using the technique of Section 11.3, formulas 11.14 and 11.15.

In this particular case $n_1 = n_2$. When the sample sizes are the same, fairly simple algebraic manipulation will show that formula 11.14 is identical to the square root of 11.8, with the correlation term 0. In any event, we can compute from Table 11.2 the sums of squares of deviations for the two groups, 61.04 and 85.66, respectively, substitute in formula 11.14, and find

$$s^2_{\bar{x}_1-\bar{x}_2} = 1.63$$

Referring back to Section 11.2, we find that this is identical to the variance found there when the correlation term, -1.177, was ignored. The square root of this result is 1.28, the standard error of the difference of means of two independent samples drawn from the same population.

Using equation 11.15, we find $t = 2.88/1.28 = 2.25$. The appropriate number of degrees of freedom is now 18 instead of 9. We choose the 1 percent level as previously. Referring to the table of t, we find that for 18 d.f., $t_{.995} = 2.88$. We recall that we rejected the hypothesis in Section 11.2 on the basis of matched groups. Here we find a contrary result, for t is not large enough to be judged significant at the 1 percent level.

In the test of *independent* groups which we have just made, we gain in degrees of freedom over the test of Section 11.2 for the matched experiment, which involved only 9 d.f. The effect is that the critical level for $\alpha = .01$ is reduced from 3.25 to 2.88. This means that with a gain in degrees of freedom, a lower t is required for significance in the independent design.

On the other hand, in the independent design, the error variance is larger than that of the matched design. As a consequence the denominator of the t in the former is 1.28, as compared with .673 in the latter. This means that the matched design yields a larger t. This may be accounted for by the relatively high correlation resulting from the matched design $r_{12} = .73$. We conclude that even though added degrees of freedom tend to make it easier to reject (or find significance) in the independent design, the greater precision (smaller error variance) in the matched design will produce a higher t if the correlation is sufficiently high. This demonstrates the major advantage of matching. The case illustrated here, however, is very unusual. Educational experiments cannot often be planned so that matched outcomes will correlate as high as .73. Unlike the present example, the gain in degrees of freedom resulting from using the independent design often is not offset by the correlation term. Only if the basis upon which subjects are paired is sufficiently related to the criterion measure will matching pay.

11.6 WHEN THE ASSUMPTIONS ARE NOT VALID

We have emphasized the assumptions of normality of distributions and equality of variances when using the t test and the pooled-variance estimate of Section 11.3. If these conditions do not hold, the use of the ratio of equation 11.15 and the t table will lead to a bias in tests of significance. What do we do if we have reason to doubt the validity of these assumptions in analyzing a given set of data? Fortunately, as theorem d of Section 7.1 states, sampling distributions of means from nonnormal populations tend to normality. In small samples it is often difficult to test the tenability of the assumption of normality of distribution. However, several investigations suggest that considerable departures from normality do not greatly invalidate the application of equations 11.14 and 11.15.

Any reason to suspect homogeneity of variance may be another disturbing element. This is particularly true if n_1 and n_2 differ greatly. Several different methods have been developed to handle this situation.

If sample sizes are unequal and it is suspected that variances are unequal, compute the statistic

$$v = (\bar{X}_1 - \bar{X}_2)/(\sqrt{s_1^2/n_1 + s_2^2/n_2} \qquad (11.17)$$

If both n_1 and n_2 are large, at least 30, consider v a normal deviate, z, and proceed accordingly to test the hypothesis. If the sample sizes are small, use the t distribution with $(n_1 + n_2 - 2)$ d.f.

The procedures of equations 11.15 and 11.17 have been compared under conditions of unequal variances. Major conclusions are:[1]

(*a*) When $n_1 = n_2$ the two ratios are identical and lead to identical statistical tests. Moreover, if $n_1 = n_2$, even when s_1^2 and s_2^2 differ widely, the type I error (probability of rejecting a hypothesis when true) differs little from the nominal values (that is, $\alpha = .05$ or $\alpha = .01$).

(*b*) When n_1 and n_2 are nearly equal, there is less bias in the application of equation 11.15.

(*c*) When n_1 and n_2 differ widely, there is less bias in the use of equation 11.17.

(*d*) For either equation 11.15 or 11.17, the chance of detecting a real difference is greater when the larger sample n is taken from the population with the larger variance.

With an example we will show how the use of v compares with the t test of Section 11.3. We assume two independent samples with:

$$\begin{array}{lll} n_1 = 10 & \bar{X}_1 = 25.3 & s_1^2 = 4 \\ n_2 = 8 & \bar{X}_2 = 22.4 & s_2^2 = 12 \end{array}$$

The observed difference between means is 2.9. The ratio of the variances is $F = 3.00$. Although reference to the F table for 7 and 9 d.f. shows us that the difference between the variances is not significant, nevertheless, the three-to-one ratio represents a considerable magnitude of discrepancy between the two variances. We substitute in equations 11.14 and 11.15 to find $t = 2.9/1.30 = 2.23$. For 16 d.f., $t = 2.12$ at the 5 percent level for a two-sided test. This test would, therefore, lead us to reject the hypothesis at the 5 percent level, and the differences would be declared significant. If we now substitute in equation 11.17 we find $v = 2.9/1.38 = 2.10$. We use the same critical value of t for this test at 16 d.f. and at the 5 percent level, that is, 2.12. According to this test we are in the acceptance region, the difference between the means would, therefore, be declared *not* significant. This example is one which illustrates precisely the occasion when distinctions such as this really become problems to the research worker.

The two methods would have led to identical action if a smaller difference had been found between the two means so that it would not have been significant on either test, or if the observed sample means had differed so much as to be significantly different on each test.

[1] Cf. Gronow, reference 4.

There are several other tests for differences between means when variances are unequal which are commonly encountered in educational research literature and which a statistician may have occasion to use. One of these is the Behrens-Fisher test. A quantity d defined the same as v, (equation 11.17) is computed. This is compared in a table with "fiducial limits" of d for various significance levels and appropriate numbers of degrees of freedom. The table is also entered with the angle whose tangent is the ratio of the standard errors of the two means.[1] We have found $d = v$ to be 2.10 in the present example. The tabled value of v at the 5 percent level would be near 2.30 for the Behrens-Fisher test. The observed value is within the interval. We would therefore *not* reject the hypothesis. The difference between means again would be considered nonsignificant. This result agrees with the test of v made by entering the t table.

Two other tests used for the case of unequal variance further confirm the inaccuracy of the t test using equation 11.15. The first of these tests has been suggested by Cochran and Cox.[2] A critical value of t at a given level is found from a table of t values corresponding to the degrees of freedom for each of the two samples. A weighted mean of these t values, weighting with respect to the variances of the means, is used as the critical t value for making the test. At the 5 percent level the critical value of t for the first group with 9 d.f. is $t_1 = 2.26$. For the second group with 7 d.f. $t_2 = 2.36$. The weighted critical value at the 5 percent point would be found by substituting in

$$t' = \frac{s^2_{\bar{x}_1}t'_1 + s^2_{\bar{x}_2}t'_2}{s^2_{\bar{x}_1} + s^2_{\bar{x}_2}} \tag{11.18}$$

The result is $t' = 2.34$. Our observed value of v, 2.10, does not reach this level. Notice that this is a more conservative test than others. A considerably larger value for v is required before we would make a decision of rejecting the hypothesis or declaring the difference statistically significant.

Another approximate method for handling the case with unequal variances has been proposed by Welch.[3] This also uses the statistic v and the critical level is found in a t table. The only difference is that Welch proposes that in using the statistic v, the t table be entered with

$$\text{d.f.} = \frac{(s^2_{\bar{x}_1} + s^2_{\bar{x}_2})^2}{\dfrac{(s^2_{\bar{x}_1})^2}{n_1 + 1} + \dfrac{(s^2_{\bar{x}_2})^2}{n_2 + 1}} - 2 \tag{11.19}$$

[1] Table V-1 in R. A. Fisher and F. Yates, reference 3, p. 52.

[2] W. G. Cochran and G. M. Cox, *Experimental Designs*, John Wiley and Sons, 1950.

[3] B. L. Welch, "The Significance of the Difference between Two Means When the Population Variances Are Unequal," *Biometrika*, 29: 350 (1938).

Substituting, we find d.f. = 3.61/.2645 − 2 = 11.6. We use the number of degrees of freedom at the nearest whole number, which in this case is 12, and find for 12 d.f. at the 5 percent level t is 2.18. Notice that this is a somewhat higher critical value of t than that found by using d.f. $= n_1 + n_2 - 2$. This, however, reconfirms our previous tests which have been recommended for the case of unequal variances. We failed to find significance in differences between the means. All the methods reverse the decision which would be made by the method of equation 11.15, which assumes homogeneity of variance.

There is an increasing number of "nonparametric" tests being developed which, though differing in efficiency, nevertheless are useful in many situations.[1] Additional alternatives include finding different measures or transforming measures. A common measure used in education is time required to complete successfully a given task. Scores on such measures often yield skewed distributions. If alternative measures of proficiency and success and performance can be found, the difficulty may disappear. Unequal variances can sometimes be rectified by using the square root of X or the logarithm of X as measures. Scores can be normalized by methods discussed in Chapter 6. Often experiments can be designed and measurements planned to avoid the difficulty of bias in the application of known tests of significance. Difficulties may be avoided, for instance, if it can be arranged to have $n_1 = n_2$, or if samples are of unequal size to have larger n in the sample with the larger variance.

11.7 COMPARING TWO PROPORTIONS

In Section 7.13 we saw that the variance of a proportion, p, in a binomial population is pq/n. Similar to equation 11.11 is the standard error of the difference between the proportions in two independent binomial populations

$$\sigma_{p_1-p_2} = \sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}} \qquad (11.20)$$

The parameters p_1 and p_2 for the two universes are not known in practical problems, but they can be estimated by p_1' and p_2'. The observed proportions in two independent samples can be substituted in equation 11.20 to compute an approximate standard error of the difference. The observed difference, $p_1' - p_2'$, can then be divided by this standard error. The result is approximately normally distributed if the samples are sufficiently large and the observed proportions are neither extremely small

[1] See references 1, 7, and 8 of this chapter, and references 17 and 20 of Chapter 12.

nor large. We may then use the normal distribution, using z as an approximate concerning differences between proportions.

A better procedure parallels and follows the logic of equations 11.12 and 11.14. The null hypothesis, $H: p_1 = p_2 = p$, is the hypothesis that the independent samples are drawn from a single universe with population value, p. The standard deviation of the difference between two independent random samples from a binomial population is

$$\sigma_{p_1-p_2} = \sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \tag{11.21}$$

In an actual problem we usually do not have the parameter, p. In applying equation 11.21 we estimate it as the proportion of the two samples combined as follows:

$$p' = \frac{n_1p_1' + n_2p_2'}{n_1 + n_2}$$

Here p' is the estimate of the common proportion from which the two samples were drawn. The estimate $q' = 1 - p'$.

The ratio of the difference between the proportions to their standard error is approximately normally distributed so that we may test our hypothesis by means of the ratio

$$z = \frac{p_1' - p_2'}{\sqrt{p'q'n/n_1n_2}} \tag{11.22}$$

where $n = n_1 + n_2$. The denominator in equation 11.22 is equivalent to the expression in equation 11.21, with the estimated parameters p' and q' substituted for the parameters p and q and with algebraic rearrangement which facilitates computation.

We illustrate this application with the hypothetical data of Table 11.3

TABLE 11.3

TWO SAMPLES FROM A BINOMIAL POPULATION

Sample	Number of Subjects		
	Having X	Not Having X	Total
I	500	500	1000
II	550	650	1200
Both	1050	1150	2200

which exhibits two samples. In sample I, $n_1 = 1{,}000$. In sample II, $n_2 = 1{,}200$. For each sample we have information concerning the number possessing some characteristic X: 550 of the 1,000 in sample I, or $p_1' = .50$ have the characteristic; in sample II, 550 of the 1,200 subjects, or $p_2' = .4583$ have the characteristic. The samples are independent. Sample I may be a random sample of registered voters in the rural part of the school district, sample II a random sample of registered voters in the urban part of a school district. The proportions may be considered as proportions having children in school, or being in favor of some proposal. We may also imagine the example as experimental in nature. Sample I might be a group of subjects under one set of experimental conditions and sample II a group of subjects under another another set of experimental conditions. The characteristic might be visualized as success in completing some type of performance test or other experimental task. It is emphasized that in either case the 1,000 individuals in the first sample are different from the 1,200 individuals in the second sample.

To test our hypothesis at the 5 percent level we shall use the two-sided criterion, $z_{.975} = 1.96$. Substituting in equation 11.22, we find that $z = .0417/.0214 = 1.95$. We thus accept the hypothesis at the 5 percent level.

It is not surprising if the reasoning of the above test sounds very much like the reasoning of applications of χ^2 in Chapter 10. As a matter of fact, the foregoing procedure is identical to a χ^2 test of a fourfold table We could use several different formulas reported in Chapter 10, but one of the simplest is formula 10.11. There is little need for correcting for continuity in a situation similar to the present example. We therefore disregard the correction for continuity in equation 10.11. With this modification of equation 10.11 we substitute and find that $\chi^2 = 3.80$. Referring to the table of χ^2 we find that a χ^2 of 3.84 is required at the 5 percent level. This test, therefore, leads also to acceptance of the hypothesis by a small margin.

Rather tedious but elementary algebra may be used to demonstrate that the square of equation 11.22 is equal to χ^2. In our example we found $z = 1.95$. Therefore, $z^2 = 3.80$, the same as the computed χ^2. The tests are identical. If small cel! frequencies are involved, a correction for continuity may be used. In that case, formula 10.11 *with the correction* is proper.

A different procedure is required when p_1' and p_2' are derived from samples which are *not independent*. Educational research frequently requires testing hypotheses concerning differences of proportions from two different observations or measurements from the *same sample* or *matched samples* of subjects. To help visualize this situation and to simplify notation, we use the following diagram of a fourfold table.

Trial I	Trial II		Totals
	X	Not X	
X	a	b	$a + b$
Not X	c	d	$c + d$
Totals	$a + c$	$b + d$	n

We shall analyze a specific example. In Table 11.4 are data from a questionnaire item in a poll of a panel of 94 persons. The 94 persons ($n = 94$) were asked before a school survey whether they preferred pupil assignments to be "varied with pupil needs," or whether they preferred uniform assignments, or were undecided. After the survey had been completed and there had been an extensive publicity campaign, through the press, radio, and public meetings, the same 94 persons were polled a second time on their attitude concerning pupil assignments in the schools. Using the schematic diagram for cell frequencies and marginal totals, we may state the proportions in which we have interest in these terms: $p_1 = (a + b)/n = 54/94 = .5745$, and $p_2 = (a + c)/n = 62/94 = .6596$. Our universe is the population of adults in the community. We wish to know if the change in responses on this item is statistically significant.

TABLE 11.4

TWO MEASURES—SAME INDIVIDUALS

Presurvey (I)	Postsurvey (II)		Total
	Varied Assignments	Other	
Varied assignments	49	5	54
Other	13	27	40
Total	62	32	94

We should carefully note the difference between this objective and another one which can be treated by the methods we have already used for two-by-two tables. Our concern at present is not a simple test of independence of Trial I and Trial II responses. If it was, we could use the test of equation

11.22, or its χ^2 equivalent, equation 10.11. We would find $z = 5.89$ and χ^2 to be 34.7, exceedingly unlikely values under the null hypothesis. This permits us to consider with merit the tendency for persons to respond in Trial II as they did in Trial I. It does not, however, enable us to judge whether there is any importance to be attached to the sample information that, as a group, the 94 persons between trials increased from a 57 percent preference to a 66 percent preference for varied assignments. The methods of Chapter 10 and the first part of this section are not appropriate for an examination of this difference because these two percentages are not from independent samples, as is clearly demonstrated by the χ^2 of 34.7. The proportions p_1 and p_2 are correlated.

The development of equation 11.10 will lead to the following formula for the standard error of the difference between two correlated proportions:

$$\sigma_{p_1-p_2} = \sqrt{\sigma^2_{p_1} + \sigma^2_{p_2} - 2\rho_{p_1p_2}\sigma_{p_1}\sigma_{p_2}} \tag{11.23}$$

By substituting the formula (13.16) for the fourfold correlation coefficient, ϕ, and expressing variances in terms of p, q, and n, it is possible to show that equation 11.23 is equivalent to the following:

$$\sigma_{p_1-p_2} = \sqrt{(b + c)/n^2} \tag{11.24}$$

Using z, the ratio of the difference to its standard error, we may test our hypothesis with the normal curve where

$$z = \frac{p_1 - p_2}{\sqrt{(b + c)/n^2}} \tag{11.25}$$

The appropriate test from the χ^2 approach may be shown to be precisely the square of equation 11.25. In simplest terms it is

$$\chi^2 = \frac{(b - c)^2}{b + c} \tag{11.26}$$

The interesting thing about the χ^2 formula is that the test is based entirely on cells b and c. These are respectively frequencies of individuals in the first trial who changed from characteristic X to Not X in the second trial, and who changed from Not X in the first trial to X in the second trial. In Table 11.4 it is the thirteen people who shifted from responses not in favor of varied assignments to responses which did favor them, and the five persons who changed from support of varied assignments to the other category with which we deal. Substituting in equation 11.26, we find that $\chi^2 = 64/18 = 3.56$. This is less than that required for 5 percent significance. Therefore, we consider tenable the hypothesis that the adults of the community were no different in their attitude toward varied assignments after the publicity program.

When this test is used with small cell frequencies it is advisable to use a correction for continuity. The formula for χ^2 may be modified to take this correction into account as follows:

$$\chi^2 = \frac{(|b - c| - 1)^2}{b + c} \tag{11.27}$$

This is the square of equation 11.25 with correction for continuity.

11.8 COMPARING TWO CORRELATIONS

The principle that was implied in equation 11.3 and in the development of 11.11 we have used several times in this chapter. It is the theorem that if two variables, X_1 and X_2, are independently and normally distributed, their difference is normally distributed with variance equal to the sum of their respective variances. This principle is involved in a test of significance of differences between correlations observed from independent groups. The z transformation explained in Section 9.5 was seen to be approximately normally distributed. Hence, if we transform the correlation coefficients observed in two samples to z', the variance of their difference will equal the sum of their variances thus:

$$\sigma^2_{z'_1 - z'_2} = \left(\frac{1}{n_1 - 3}\right) + \left(\frac{1}{n_2 - 3}\right) \tag{11.28}$$

To test the hypothesis H : $\rho_1 = \rho_2$, that there is no difference in the correlation in the two populations, we use

$$z = (z'_1 - z'_2)/\sqrt{1/(n_1 - 3) + 1/(n_2 - 3)} \tag{11.29}$$

Two different reading tests, purporting to measure two entirely different types of reading skill, were administered to a group of 18 "poor" readers and to a second group of 21 "good" readers. The observed correlations were .69 and .43, respectively. The transformation for $r_1 = .69$ is $z'_1 = .848$, and for $r_2 = .43$ is $z'_2 = .460$. The sample sizes are $n_1 = 18$ and $n_1 = 21$. We find that

$$z = .388/\sqrt{.06667 + .05556} = .388/.350 = 1.11$$

Since this falls short of significance, the hypothesis that the samples are from a population with common ρ is tenable.

In the foregoing we were dealing with *independent* groups in comparing two correlations of the same two variables X and Y. An entirely different problem exists when we have *two correlations obtained from the same subjects.* Such correlations would not be independent. A special case of this situation concerns correlations obtained from the same single group of subjects, between each of *two* separate measures and a third.

In an analytical study of problem solving, three measures were approximately defined as follows: X_1 = success in problem solving, X_2 = a measure of skill in interpreting data, and X_3 = a measure of persistence in pursuing hypotheses. In a sample of forty subjects the correlation between success in problem solving and the interpretation measure was found to be .47, between problem solving and the persistence measure, .64. It is desired to test the significance of the difference between these two correlations. The following may be used to test the hypothesis of no difference:

$$F = \frac{(r_{12} - r_{13})^2\,(n-3)(1 + r_{23})}{2(1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2r_{12}r_{13}r_{23})};\quad \text{d.f.} = (1), (n-3) \qquad (11.30)$$

In the example given, the added information that $r_{23} = .53$ is required. We may then substitute to find

$$F = \frac{(.0289)(37)(1.53)}{2(1 - .221 - .410 - .281 + .319)} = \frac{1.64}{.814} = 2.01$$

The test of significance is completed by looking up critical values of F for 1 and $n - 3 = 37$ d.f. We find that the result is not significant. When the numerator of F contains 1 d.f. it is the square of t with degrees of freedom for the denominator. A t table can thus be used for this test. We take the square root of the observed F and find $t = 1.42$, for 37 d.f. This is not significant.

EXERCISES

1. The Standard deviation of part I of a test is 16 and the standard deviation of part II is 25. If the correlation between the two parts is .75, what is the standard deviation of a total score made up of the sum of part I and part II scores?

2. State the assumptions that are made when the t distribution is used in testing a hypothesis concerning the difference between two means.

3. Under what circumstances is it advantageous to design an experiment by "matching" or pairing observations in two groups?

4. Assume that the only information available to you from Appendix G is that concerning the physical science test scores of students sampled from high schools A and D. Assume furthermore that the two samples are from infinitely large populations. Test the hypothesis of no difference in means at the .01 level. What are the assumptions underlying this test? What assurance have you that these conditions have been met? Make such tests as you can to support your conclusion concerning assumptions.

5. The mean and standard deviation of measures from a sample of 17 pupils are, respectively, 125 and 10. A second sample consists of 29 pupils with a mean of 114 and a standard deviation of 25.

(*a*) Test the hypothesis that the two samples are from populations with the same variances.

(*b*) Assuming that the two population variances are not the same, select an appropriate test for the hypothesis that there is no difference between the means of the two populations. (Use the .05 level.)

(*c*) What other tests might have been used in this case?

(*d*) Why is this an approximate test?

(*e*) Is the real probability of a Type I error equal to .05, greater than .05, or less than .05?

6. Test the significance of the difference at the .01 level between the correlation coefficients, r_{xy}, for high schools A and D of Appendix G. State the hypothesis to be tested and interpret your results.

7. In a study of the relationship of auditory discrimination to articulatory defects of elementary school children, thirty normal-speaking elementary children were matched with thirty children diagnosed as having functional articulatory defects. The matching was on the basis of age, sex, grade, and intelligence. A test of speech-sound discrimination was given to the two groups. For the defective group the mean was found to be 28.83 and the standard deviation 4.10. The mean for the control group was 12.37 and the standard deviation was 3.37. A t ratio of 23.56 was found.

(*a*) What evidence is there that the investigator used the proper formula for t for matched groups?

(*b*) Assuming that the computation is proper, test the hypothesis at the 1 percent level, using t with the appropriate number of degrees of freedom.

(*c*) How would you compute the correlation between pairs from the data given?

(*d*) Was there much advantage to the investigator in matching?

8. In an experiment in remedial instruction in arithmetic, a population of pupils was classified in eight groups comparable in mental age and arithmetic achievement test score. A pair of subjects was randomly selected from each of the eight categories. One of the two subjects of each pair was selected at random for assignment to a remedial instruction group. The other member of each pair was assigned to the control group. At the conclusion of the experimental instruction scores on a criterion measure were as follows:

Group	Pair							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Control	8	7	10	9	8	6	9	7
Remedial	10	8	11	10	14	10	12	8

Test the hypothesis of no difference in criterion scores between the two groups specifying the hypothesis you are testing, the level of significance chosen, and the assumptions made. In one sentence state your conclusion. Analyze the data so as to show the contribution of the "matching" feature of the experiment. What comments may be made on experimental design?

9. In the published results of testing a group of elementary pupils it is reported that the variance of test I was 16, the variance of test II was 9, and the variance of differences in score on the two tests was 7. By means of equation 11.2 find the correlation between the tests.

10. Repeat the test of Ex. 4, Chapter 10, using the methods of this chapter and compare results. In what respects would your test differ if the hypothesis was that the proportion of boys intending college was greater than the proportion of girls intending college?

11. An experiment was conducted to study the effects of an educational film on attitudes of pupils. A test was given to a sample of 60 pupils before the showing of the film and repeated after the showing of the film. One item of the test concerned the attitude of pupils on an issue. Pupils responded by indicating whether they favored or opposed a line of action. The distribution of the responses was as follows:

Pre-test	Post-test		Total
	Favor	Oppose	
Favor	15	17	32
Oppose	19	9	28
Total	34	26	60

State and test two hypotheses which would be of interest in analyzing these data.

12. Measures of "group cohesion" were computed for each of two randomly assigned groups of pupils in a classroom following a competitive group experiment. The two groups were randomly assigned a motivational "treatment" before the experiment. The treatment was intended to produce widely divergent group goals. The experiment was repeated in 25 classrooms. In 8 of the 25 experiments the group cohesion measure for treatment A exceeded that for treatment B. In the remaining 17 experiments the groups with the B treatment exceeded the groups with the A treatment. At the .01 level, test the hypothesis that the two groups are from the same population? State specifically the hypothesis to be tested. In this test what assumptions are made regarding the distribution of "group cohesion" measures? Of differences between groups on this measure? Why is this type of test sometimes called a "distribution-free" test? Why sometimes a nonparametric test?

13. Define the two types of error in testing a statistical hypothesis. Give examples of situations in which an educational statistician may need to take both of them into account.

14. What risk does the experimenter take if his statistical test of a null hypothesis is low in power? In general, how should the experimenter increase the power of his test, by establishing a low value for α or by establishing a high value for α? What are some of the ways of increasing the power of a test of "the significance of differences" between two populations? From the standpoint of how best to advance human knowledge, would you consider it the best policy for an experimenter always to proceed conservatively in the sense of choosing very small values for α such as .01 or .001?

15. An experiment has been conducted to test the usefulness of a new method of teaching arithmetic in the third grade. Ten subjects were assigned to group I, an experimental group, and independently ten to a control group, Group II. Assume

that the variance of the criterion measure is known to be the same for both groups and equal to 25. It is then appropriate to use equation 11.12 for the standard error of the difference between means and the normal distribution in testing the one-sided hypothesis, $H : \mu_1 - \mu_2 \leq 0$, that the new method is not superior to the conventional one. Suppose the test is to be made at the 5 percent level.

(*a*) What is the critical region for testing the hypothesis?

(*b*) Suppose it is found that $\bar{X}_1 - \bar{X}_2 = 3.35$, would the hypothesis be accepted or rejected?

(*c*) What would be the probability of accepting the hypothesis when in fact $\mu_1 - \mu_2 = 2.0$? This is the probability of what type of error?

(*d*) What would be the power of this test in detecting a difference in favor of the experimental method by as much as 2.0?

(*e*) Suppose that the new method would result in considerable economies in teaching and greatly improved educational advantages if $\mu_1 - \mu_2$ was *at least* 2.0. Suppose, furthermore, that it was thus considered important to design an experiment so that there would be at least an 80 percent assurance that a true difference of as much as 2.0 would be detected. How many subjects would need to be assigned each group to test the hypothesis $H : \mu_1 - \mu_2 \leq 0$, with $\alpha = .05$, so that if $\mu_1 - \mu_2 \geq 2.0$ then $(1 - \beta)$ would be at least .80?

REFERENCES

1. Dixon, Wilfrid J., and Frank J. Massey, *Introduction to Statistical Analysis*, New York, McGraw-Hill Book Co., 1951, Chapter 17.
2. Edwards, Allen L., *Experimental Design in Psychological Research*, New York, Rinehart and Co., 1951, Chapter 8.
3. Fisher, Ronald A., and Frank Yates, *Statistical Tables for Biological, Agricultural and Medical Research*, Fourth Ed., New York, Hafner Publishing Co., 1953.
4. Gronow, D. G. C., "Test for the Significance of the Difference Between Means in Two Normal Populations Having Unequal Variances," *Biometrika*, 38 : 252-56, June 1951.
5. Johnson, Palmer O., *Statistical Methods in Research*, New York, Prentice-Hall, 1949, Chapter 5.
6. Merrington, Maxine, and Catherine M. Thompson, "Tables of Percentage Points of the Inverted Beta Distribution," *Biometrika*, 33: 73-88, April 1943.
7. Mood, Alexander M., *Introduction to the Theory of Statistics*, New York, McGraw-Hill Book Co., 1950, Chapter 16.
8. Walker, Helen M., and Joseph Lev, *Statistical Inference*, New York, Henry Holt and Co., 1953, Chapters 7 and 8.
9. Welch, Bernard L., "The Significance of the Difference between Two Means When the Population Variances are Unequal," *Biometrika*, 29: 350-62, February 1938.
10. Wert, James E., Charles O. Neidt, and J. Stanley Ahmann, *Statistical Methods in Educational and Psychological Research*, New York, Appleton-Century-Crofts, 1954, Chapter 8.

CHAPTER 12

Comparison of More Than Two Groups

We have already had some experience comparing more than two groups by means of χ^2, but this comparison was limited to hypotheses concerning frequencies and proportions. It is often desired to test hypotheses concerning the means of several groups. This involves the theory of Chapter 7 and requires an extension of the ideas of Chapter 11 which are concerned with the comparison of two means. A good preparation for the study of the present chapter is a review of the major principles of Chapters 7 and 11.

12.1 A MODEL FOR TESTING THE VARIATION OF SEVERAL SAMPLE MEANS

The strategy in testing the significance of variation (over-all differences) of several sample means is to determine the feasibility of the several samples having been drawn from a single population. When we are presented with k sample means, we are interested in testing the hypothesis that the various population means are actually the same. In other words, $H : \mu_j = \mu$. Here μ_j is the mean of any one of the subpopulations (the jth subpopulation), and μ is the mean of the general population,

$$\mu = \frac{\sum\limits_{j=1}^{k} n_j \mu_j}{\sum\limits_{j=1}^{k} n_j},$$

where n_j is the number of observations in the jth subpopulation, and k is the number of such subpopulations. In short, we shall assume independent random samples from a normal population.

As an introduction to the usefulness of this model in comparing means of groups, we shall consider samples of equal size, n_j, although this, as we shall see later, is not a necessary limitation. Let us begin with a numerical illustration.

In Table 12.1 are seven samples drawn independently and at random from a normal population. The number of observations or scores in each sample is 5. Our interest lies in the first two lines of the table beneath the sample scores. The first row is the sum of the scores. We shall introduce a new symbol, T_j, to represent the sum of the sample scores in the jth group, that is,

TABLE 12.1

SEVEN RANDOM SAMPLES OF SIZE FIVE FROM A NORMAL POPULATION

	Sample						
Item	I	II	III	IV	V	VI	VII
Scores	6 1 1 2 6	5 6 2 5 5	5 3 5 6 3	2 1 8 7 4	3 5 4 3 7	8 8 6 5 4	5 4 4 5 5
T_j	16	23	22	22	22	31	23
$\bar{X}_j$	3.20	4.60	4.40	4.40	4.40	6.20	4.60
ΣX_{ij}^2	78	115	104	134	108	205	107
T_j^2/n	51.2	105.8	96.8	96.8	96.8	192.2	105.8
$\Sigma X_{ij}^2 - T_j^2/n$	26.8	9.2	7.2	37.2	11.2	12.8	1.2

$$T_j = \sum_{i=1}^{n_j} X_{ij},$$

where X_{ij} is the ith individual in the jth sample group. Since $n_j = 5$, we divide each of the sample totals by 5 to get the means designated by $\bar{X}_j$, in the second row.

These sample means, 3.20, 4.60, etc., are sample means of size 5 drawn by *random sampling* from a *normal* population. Our object is to discover whether or not they differ unreasonably under the hypothesis. We might make all possible comparisons a pair at a time. This would not only be cumbersome, but would also be unsuitable. Suppose for instance, that all pairs of means were tested by the method of equation 11.15 at the 5 percent level. By the very nature of our definition of $\alpha = .05$, we expect a significant result once in 20 tests even if the hypothesis is true. Furthermore, the t tests would not be independent and there would tend to be significant t's more frequently than indicated by the significance level chosen. Also, the nature of the hypothesis is such that a single composite test is desired. By the methods to be described we will gain in the power of our test by taking advantage of the information from all groups combined in estimating the variance of the population.

We shall use the symbol, N, to represent the number of scores in all samples combined. In general, $N = \sum_j n_j$ (read N equals summation over all j groups of the number of observations n_j). In Table 12.1 the sample sizes are the same. In this case we drop the subscript for n and $N = kn$. For the seven samples of size 5 there are thus 35 observations, $N = 35$. The mean of the jth group may be computed from $\bar{X}_j = T_j/n$.

The general mean may be found by summing the totals over all seven samples and dividing by the grand total, thus

$$\bar{X} = \sum_j T_j/N = 159/35 = 4.54$$

This result is our best estimate of the general population mean μ.

The variance of the sample means would cast doubt upon the claim made above that these are random samples from the same universe, except that some variation among them is to be expected. The size of the variance of these means certainly should be strategic in any measure or test of the reasonableness of the claim that these seven samples have been drawn at random from the same population. It should therefore be of primary interest to compute this variance.

We may express each sample mean as a deviation from the general mean, $(\bar{X}_j - \bar{X})$. Squaring these, summing them, and dividing by the appropriate degrees of freedom, we have an estimate of the variance of a population of sample means. The first sample deviates -1.34 from the general mean, the second sample mean deviates $+0.06$ from the general sample mean, and so on. The sum of squares of all seven deviations is 4.62. There is a total of seven means involved in the computation, and 1 degree of freedom is lost for computing the general mean. Therefore the appropriate denominator for our variance is $(k - 1) = 6$.

In summary, we may compute the variance of sample means as follows:

$$s^2_{\bar{x}(1)} = \sum_j (\bar{X}_j - \bar{X})^2/(k - 1) = 4.62/6 = .77$$

The subscript on the variance of the means contains "(1)" to identify this estimate, since we shall compute a second estimate later by another method.

If we can devise some reference value for purposes of determining whether or not the observed variance of means is "too large" under the hypothesis, we have solved the problem of testing the hypothesis. We now find such reference value.

A feature of Chapter 7 was the method of estimating the sampling variance of the distribution of sample means from a single sample. From equation 7.4 we know that

$$s_{\bar{x}}^2 = s^2/n$$

where s^2 is the sample estimate of the universe variance, and n is the number in the sample. For each one of the seven samples we could use equation 4.14 to compute the sum of squares so that we could compute s^2 using equation 7.3.[1] The method of equation 4.14 is to obtain the sum of squares of the *gross* scores in a sample and subtract from that the ratio of the square of the sum of the scores to the number of observations in the sample. This is shown for each of the seven samples as the last line of Table 12.1. The figures in the last line of this table, therefore, represent for each sample the sum of the squares of *deviations* from the corresponding sample mean. Since we have seven samples, it is to our advantage to make use of the information from all seven of them combined. Rather than applying equation 7.4 on the basis of only one of the samples, we should have a better estimate of the variance of sample means if we pool or combine the information from each of the seven separate samples. One method would be to compute a variance, $s_{\bar{x}}^2$, for each one and to find the average.

We shall use a slightly different numerical method leading to the same result, one which follows the scheme which will be used throughout this chapter. This is the method of pooling the sums of squares of deviations "within groups," the last line of Table 12.1. We shall call this the *within* sum of squares because, even though summed across all samples, it is based on deviations from respective sample means. This we shall use as a numerator for our estimate of the universe variance, s^2. The denominator will be the sum of the degrees of freedom for the k groups. In symbols this is

$$s^2 = \frac{\sum_j \sum_i (X_{ij} - \bar{X}_j)^2}{k(n-1)} = \frac{\sum_j \left[\sum_i X_{ij}^2 - T_j^2/n\right]}{k(n-1)} \qquad (12.1)$$

[1] In the remainder of this chapter the term *sum of squares* or the term *sums of squares*, standing alone, will be used to mean sum or sums of squares of *deviations* from a mean except where it is clear from the context that reference is to *gross* scores.

The first expression best describes the variance; the second is more suitable for computation.

We can find the numerator simply by summing the entries in the last row of Table 12.1. This sum turns out to be 105.6. Now for each sum of squares for each sample there is $(n - 1)$ or, in this case 4, degrees of freedom. For all k samples the total number of degrees of freedom is therefore $(7)(4) = 28$. From equation 12.1 we find that

$$s^2 = 105.6/28 = 3.77$$

From the relationship, $s_{\bar{x}}^2 = s^2/n$, we may use this estimate of the universe variance (based on the pooled sums of squares within samples) to find an estimate of the variance of means of size n to be expected under the conditions of repeated random sampling from this normal population. The result is $s_{\bar{x}(2)}^2 = 3.77/5 = .75$.

We are now ready to investigate the reasonableness of the variation of the sample means in Table 12.1. By hypothesis, these samples are random samples from a common population. In other words, the expected value of each sample mean is the same, the population mean μ, that is, $E(\bar{X}_j) = \mu$. The *actual* distribution of means yields us a variance, $s_{\bar{x}(1)}^2 = .77$. By securing the best estimate of the variance of this population and from it computing $s_{\bar{x}(2)}^2$, we obtained another estimate, independent of the *actual* variance of the means because it uses only deviations from those means. This second variance is a measure of the dispersion of these means that we would expect under the hypothesis of random sampling from a single population. Thus it is a yardstick for judging the former.

We find that the first actual variance of means, .77, is larger than .75, the value we expect. Is the discrepancy large enough to cast real doubt on the hypothesis, that variations among the observed means are attributable only to chance? We need a method of determining how much larger the actual variance must be to raise a serious question about the hypothesis that all the sample means come from the same population.

In Section 11.4 a distribution was discussed which enables us to answer such questions. The ratio of two independent variances from a normal population has a known distribution, F. Therefore, we compute the ratio of the variance based on the actual means to the variance expected under the hypothesis, and, if this ratio is considerably larger than 1.00, we would suspect the hypothesis. In the present case the ratio of the two variances is $F = 1.03$, based upon $(k - 1) = 6$ and $k(n - 1) = 28$ d.f., respectively. We can find, say, the 5 percent rejection region for testing the hypothesis, by referring to the table in Appendix I. If our observed F is equal to or exceeds the critical value we would reject the hypothesis.

As a matter of fact, the seven sets of sample values actually were drawn in a random manner from a normal population and so there was one chance in twenty that we would find an F-ratio high enough to cause rejection at the 5 percent level.

12.2 PARTITIONING THE SUM OF SQUARES

The basic idea of the process which we have just completed is the relationship of the sums of squares used in the analysis. We shall rearrange our data somewhat in order to see this relationship clearly.

We are considering the case of k groups (samples) of n individual scores each. There are three deviations which may be computed for each individual:

(1) The deviation of the individual score from the *general* mean, $(X_{ij} - \bar{X})$.

(2) The deviation of the individual score from the mean of the group (or sample) to which it belongs, $(X_{ij} - \bar{X}_j)$.

(3) The deviation of the individual's group mean from the general mean, $(\bar{X}_j - \bar{X})$.

It is readily seen that these three deviations are related in the following manner:

$$(X_{ij} - \bar{X}) = (X_{ij} - \bar{X}_j) + (\bar{X}_j - \bar{X})$$

For the first score, 6, in the first sample of Table 12.1, this is verified as $(6 - 4.54) = (6 - 3.20) + (3.20 - 4.54)$ or, $1.46 = 2.80 - 1.34$. The same relationship holds for each X_{ij} in the table.

We may now square and sum over all i in the jth group to find

$$\begin{aligned}\sum(X_{ij} - \bar{X})^2 &= \sum_i[(X_{ij} - \bar{X}_j) + (\bar{X}_j - \bar{X})]^2 \\ &= \sum_i(X_{ij} - \bar{X}_j)^2 + \sum_i(\bar{X}_j - \bar{X})^2 + 2(\bar{X}_j - \bar{X})\sum_i(X_{ij} - \bar{X}_j)\end{aligned}$$

The second term on the right side of this equation involves the summation of a constant, $(\bar{X}_j - \bar{X})^2$, for all n members of the group, and becomes $n(\bar{X}_j - \bar{X})^2$. The last term vanishes since, for the jth group (that is, any group), the sum of the deviations from the group mean is zero. That is, $\Sigma(X_{ij} - \bar{X}_j) = 0$.

Hence, in the jth group,

$$\sum_i(X_{ij} - \bar{X})^2 = \sum_i(X_{ij} - \bar{X}_j)^2 + n(\bar{X}_j - \bar{X})^2 \tag{12.2}$$

It is worth while examining this relationship with reference to what we observed in Chapter 4. We noted there that the mean-square deviation from a group mean was less than the mean-square deviation from some other value. At the present moment we are not dealing with *mean*-square deviations, but the same principle is involved since equation 12.2 shows us that the sum of squares about $\bar{X}$ is greater than the sums of squares about the group mean, $\bar{X}_j$, by an amount equal to n times the square of the difference. In this analysis $\bar{X}$, relative to the jth group, corresponds to X_a (some value other than the group mean) in the notation of Chapter 4. To repeat, the *within* group sum of squares, $\sum_i (X_{ij} - \bar{X}_j)^2$, is *less* than the sum of squares taken about the general mean, or for that matter any value other than $\bar{X}_j$.

Suppose now that we sum over all groups. Then,

$$\sum_j \sum_i (X_{ij} - \bar{X})^2 = \sum_j \sum_i (X_{ij} - \bar{X}_j)^2 + n \sum_j (\bar{X}_j - \bar{X})^2 \qquad (12.3)$$

The left side of equation 12.3 is the grand sum of squares of all nk individuals taken about the general mean. We will call it the *total sum of squares*. The first term on the right is the *pooled* sum of squares within groups around their respective group means. It is the same pooled sum of squares used in equation 12.1, the *within* sum of squares within groups around their respective group means. It is the same pooled sum of squares used in equation 12.1, the *within* sum of squares. The last term is the sum of squares among means. We will call it the *between* sum of squares, since it represents the variation among means of groups. The presence of the factor n indicates that this squared deviation component is representative of all n individuals in each of the k groups. Therefore, each of the $N = nk$ individuals is represented in each of the three sums of squares in equation 12.3.

Equation 12.3 may now be stated verbally;

$$\begin{array}{ccccc} \textit{total} & & \textit{within} & & \textit{between} \\ \text{sum of} & = & \text{sum of} & + & \text{sum of} \\ \text{squares} & & \text{squares} & & \text{squares} \end{array}$$

This characteristic of sums of squares of components adding to the total is of major importance in the method of analysis of variance which we are about to describe. The principle is applicable to more complex problems of comparing groups.

12.3 THE ANALYSIS OF VARIANCE

If we now divide each sum of squares by the appropriate number of degrees of freedom, we may express results in terms of variance or *mean square*. For the total, we lose 1 d.f. because of $\bar{X}$. Therefore, the degrees of freedom for *total* is $(N - 1)$ or, since groups are of equal size, $(nk - 1)$. The degrees of freedom corresponding to the *within* sum of squares was already shown in equation 12.1 to be $k(n - 1)$. One degree of freedom is lost for each $\bar{X}_j$ leaving $(n - 1)$ degrees of freedom for each group, or $k(n - 1)$ for the k groups. In computing the *between* sum of squares, 1 d.f. is lost in computing $\bar{X}$. The sum of squares *between* reflects all N individuals, but only the k group means are required to calculate it. Therefore, the corresponding number of degrees of freedom is $(k - 1)$.

Since

$$\underset{total}{(nk - 1)} = \underset{within}{k(n - 1)} + \underset{between}{(k - 1)} \tag{12.4}$$

we find that the degrees of freedom partition into components just as do the sums of squares.

The sums of squares of equation 12.3 and the corresponding degrees of freedom of 12.4 with the resulting variances are displayed in Table 12.2.

TABLE 12.2

ANALYSIS OF VARIANCE—k SAMPLES OF EQUAL SIZE n

Source of Variation	Sum of Squares	Degrees of Freedom	Variance
Total	$\sum_j \sum_i (X_{ij} - \bar{X})^2$	$nk - 1$	$\dfrac{\sum_j \sum_i (X_{ij} - \bar{X})^2}{nk - 1}$
Between groups*	$n\sum_j (\bar{X}_j - \bar{X})^2$	$k - 1$	$\dfrac{n\sum_j (\bar{X}_j - \bar{X})^2}{k - 1}$
Within groups	$\sum_i \sum_j (X_{ij} - \bar{X}_j)^2$	$k(n - 1)$	$\dfrac{\sum_j \sum_i (X_{ij} - \bar{X}_j)^2}{k(n -1)}$

* More grammatically, *among* groups since there are usually more than two. Both designations are used in analysis of variance.

The table is called an *analysis of variance* table. The construction of such a table and the ensuing tests of significance are termed *analysis of variance.* It is of interest to note that the table is appropriately named since it does contain variances, although the process as a whole might better be termed *analysis of means* for our objective is the comparison of means among groups.

Following the scheme of Table 12.2, we construct an analysis of variance table for the data of Table 12.1. Instead of operations as defined in Table 12.2 for finding sums of squares (ss), when samples are of equal size n we will follow a much more efficient scheme of computation as follows:

$$\text{ss total} = \sum_j \sum_i X_{ij}^2 - T^2/N$$

$$\text{ss between} = \sum_j (T_j^2/n) - T_j^2/N$$

where T is the grand total of all scores, that is, $T = \sum_j T_j$.

$$\text{ss within} = \sum_j \sum_i X_{ij}^2 - \sum_j (T_j^2/n)$$

To find the sum of squares for total, we sum the totals of the squares of raw scores for all individuals in all groups and subtract the ratio of the square of the grand total of raw scores to the grand total number of individuals. For the other two sums of squares we also need the sum over all groups of the ratio of the square of the group total to the number of individuals in the group. From the above relationships it is a simple matter to finish our computations. For Table 12.1 we compute

$$\begin{aligned} \text{ss total} &= 851.00 - 722.31 = 128.69 \\ \text{ss between} &= 745.40 - 722.31 = 23.09 \\ \text{ss within} &= 851.00 - 745.40 = 105.60 \end{aligned}$$

We ordinarily do not compute the *total variance*, as it is of no use to us in the analysis we are about to make. The total variance is independent of neither of the other two variances so that we may not appropriately make use of variance ratios with it. On the other hand, the variance *between* and the variance *within* are of considerable importance. They are independent and may be compared by means of the variance ratio, F. The ratio of s_b^2, the *between* variance to s_w^2, the *within* variance is the statistic

$$F = s_b^2/s_w^2 \tag{12.5}$$

For Table 12.3, $F = 3.85/3.77 = 1.02$. This agrees with the F obtained in Section 12.1 except for rounding errors, an outcome which may be explained by a comparison of $s_{\bar{x}(1)}^2$, $s_{\bar{x}(2)}^2$, s_b^2, and s_w^2. Note that $s_b^2 = ns_{\bar{x}(1)}^2$ and $s_w^2 = ns_{\bar{x}(2)}^2$. The *within* variance, s_w^2, is identical to the value computed for equation 12.1 to estimate the variance of the distribution. In

Section 12.1 we divided this estimate by $n = 5$ in order to obtain an estimate of the variance of *sample means* which might be attributed to random sampling. This was done so that we could make a direct comparison with $s^2_{\bar{x}(1)}$, the variance of sample means. In this section, the numerator of the F ratio from equation 12.5 is n times the variance of the means. From the relation $s^2 = ns^2_{\bar{x}}$, we thus have in Tables 12.2 and 12.3 a *between* variance which is an estimate of the *population* variance—not a variance of means—even though it is derived from a sum of squares of deviations of means.

TABLE 12.3

Analysis of Variance of Data in Table 12.1

Source of Variation	Degrees of Freedom	Sum of Squares	Variance
Total	34	128.69	—
Between samples	6	23.09	3.85
Within samples	28	105.60	3.77

A somewhat modified interpretation is required and is standard in analysis of variance problems. It hinges upon the meaning of the numerator and denominator and the F ratio used in the present section.

In this section the numerator is an estimate of *the variance, σ^2, of a population which would produce sample means as different as those observed on the assumption of random sampling.* The denominator, based on the within variance, is an independent estimate of the variance, σ^2, of the population from which each sample is assumed to be drawn. It is assumed that the group variances are the same. The test of our hypothesis is a test of the *between variance* against the *within variance.* If the ratio of these two mean squares is sufficiently great, it leads us to reject the hypothesis of no difference in the means.

In using the F distribution we are usually interested in a rejection region in only the right-hand tail, since only to the extent that the ratio is greater than 1 does it indicate inequality of the means from which the samples were drawn. Examining the table in Appendix I, we see that for 6 and 28 d.f., the critical value of F is 2.44 at the 5 percent level and 3.53 at the 1 percent level.

The previous discussion has been purely for the purpose of developing the basic ideas used in problems requiring the testing of significance of the differences among several sample means. We purposely illustrated with an example in Table 12.1 designed to show what is likely to happen when the hypothesis is true, as indeed it was. The samples in Table 12.1 were actually as stated, samples drawn from a single normal population. Thus it is known that the differences among the sample means are due only to the process of random sampling from a normal population. The result of our statistical test is thus no surprise to us.

In the real situation of an experiment, the "effect" of different conditions under which samples may be drawn is not known in advance. It is the purpose of the statistical analysis which we have described to throw light on the possible presence or absence of such "effects." We shall now examine features of the model when the null hypothesis is *not true*.

12.4 WHEN THE NULL HYPOTHESIS IS FALSE

In an actual situation our data may consist of k samples of size n, each from k populations with unknown means $\mu_1, \mu_2, \ldots, \mu_k$. We assume that, though the means differ, the populations have a common variance, σ^2. Our k samples are selected in such a manner as to represent populations differing on the basis of some factor. We wish to examine the "effect" of this factor on variable X_{ij}. The groups may be distinguishable either on the basis of (*a*) variations in some *treatment* given the individuals in them or (*b*) some *environmental* characteristic already known about them. The experimenter creates the former by some type of manipulation, whereas group differences in the latter already exist. The methods we are discussing apply to either situation, provided that the samples in each group are random and the assumptions of normality and homogeneity of variance are tenable.

Examples of group comparisons based upon experimental variations in *treatment* are:

Basis of Grouping	Dependent Variable (X_{ij})
Variation in type of fertilizer used.	Yield of crop.
Variation in method of instruction.	Achievement in subject.
Variation in role of group leader.	Morale of group members or effectiveness in achieving group task.
Variation in elapsed time from learning to testing.	Retention of items studied.
Variation in size of type of reading matter.	Some measure of reading or readability.

Examples of group comparisons based on *environmental* factors or known characteristics of individuals are:

Basis of Grouping	Dependent Variable (X_{ij})
Breed of animal.	Gain in weight under some standard treatment.
Shape of school building design.	Construction cost per cubic foot.
Type of high school from which graduated.	College grade point average.
Occupation of father.	Score on vocational interest inventory.
Geographic region.	Average annual expenditure per pupil for heating school plant.

Whether we are interested in the "effect" of an experimental treatment, or the "effect" of a known characteristic, we define *effect* of the factor under consideration upon the dependent variable, X_{ij}, as $\mu_j - \mu$. The test of the previous section is equivalent to the test of the hypothesis that the effect, $\mu_j - \mu$, for all groups is zero.

It is informative to see why it is that we would expect the F ratio to be greater than 1.00 when the effects are *not* all zero. As in the previous section, the *within* variance, s_w^2, is an unbiased estimate of the common variance σ^2, that is, $E(s_w^2) = \sigma^2$. We noted also that $E(s_b^2) = \sigma^2$ when the null hypothesis is *true*. This is not the case, however, when the hypothesis is false, that is, when there are real treatment effects. When the null hypothesis is not true, the *between* variance consists of two components of variation. In this case the expected value of the *between* variance is

$$E(s_b^2) = \sigma^2 + \frac{\sum\limits_j n_j(\mu_j - \mu)^2}{k - 1} \tag{12.6}$$

In other words, if the population means differ, the *between* variance used in the numerator of the F test is an estimate not only of the variance, σ^2, but, in addition, of a second component which measures the variation among the unknown group means, μ_j. Because of this second component in equation 12.6, the greater the effects $(\mu_j - \mu)$, the greater the expected value of s_b^2, the greater the numerator of $F = s_b^2/s_w^2$, and the greater the chance of rejecting the hypothesis.

The second component of the expected value in equation 12.6 represents a variance and may be written $n_j\sigma_a^2$ for factor A if the groups of factor A represent a random sample from a population of such groups. The groupings might be, for example, "high school from which graduated." An experiment consisting of a random sample of n measures of graduates from each of k high schools taken randomly from a population of high schools, leads to one type of conclusion. An experiment consisting of

n measures of graduates randomly sampled from each of k *particular* high schools leads to another.

12.5 SAMPLES OF UNEQUAL SIZE

We will now apply the foregoing theory to actual data. In the theoretical example of Table 12.1 we used samples of equal size. Although most of the above discussion has been in those terms, the theory and general principles apply when samples are of *unequal* size. Analysis of variance with varying n is the more general case, and the procedures outlined in Table 12.4 apply with constant n also. The variance, or mean square, column contains simply the sum of squares divided by the appropriate number of degrees of freedom, and such a column always appears in a finished analysis of variance table. A comparison of Tables 12.2 and 12.4 should be made and the reasons for differences in them noticed. There are advantages, particularly in more complex analysis of variance designs, in planning experiments such that sample sizes are equal. This is sometimes difficult to accomplish in the analysis of data obtained from nonexperimental conditions. In general, arranging samples of equal size improves the power of the test and usually greatly simplifies the arithmetic.

TABLE 12.4

ANALYSIS OF VARIANCE—k SAMPLES OF UNEQUAL SIZE

Source of Variation	Sum of Squares		Degrees of Freedom
	As Defined	As Computed	
Total	$\sum_j \sum_i (X_{ij} - \bar{X})^2$	$\sum_j \sum_i X_{ij}^2 - T^2/N$	$N - 1$
Between groups	$\sum_j n_j(\bar{X}_j - \bar{X})^2$	$\sum_j (T_j^2/n_j) - T^2/N$	$k - 1$
Within groups*	$\sum_j \sum_i (X_{ij} - \bar{X}_j)^2$	$\sum_j \sum_i X_{ij}^2 - \sum_j (T_j^2/n_j)$	$N - k$

* Usually computed as remainder by subtracting *between sums of squares* or *between degrees of freedom* from corresponding value for total.

In Table 12.5 are gains in scores on the Minnesota Teacher Attitude Inventory for four groups of students in four different programs of teacher training in the senior year.[1] Can the mean gain be considered the same

[1] Adapted from Carl W. Proehl, *An Experimental Study of the Effects of Two Patterns of Professional Education in the Preparation of Secondary School Teachers* (Doctoral Thesis), Urbana, University of Illinois, 1953, pp. 143-146.

for the four "treatments"? We could compute the observed means by finding $\Sigma X_{ij}/n_j$ in each case. Such computation is unnecessary for analysis of variance, but would show that the gains vary considerably among the groups, from a small average loss for group IV to a mean gain of 23.6 for group III. With the information from Table 12.5 and the computing outline in Table 12.4, we find the sums of squares for *total* and *between* to be

TABLE 12.5

SUMMARY OF DATA ON GAINS IN MTAI SCORES FOR FIFTY SENIORS IN FOUR TREATMENTS IN A TEACHERS' COLLEGE

Treatment	n_j	ΣX_{ij}	ΣX_{ij}^2
I	9	58	1,154
II	18	79	6,097
III	11	260	10,260
IV	12	-5	1,991
All groups	50	392	19,502

$$\text{ss } total = 19{,}502 - (392)^2/50 = 16{,}429$$

$$\text{ss } between = [(58)^2/9 + (79)^2/18 + (260)^2/11 + (-5)^2/12] - (392)^2/50 = 3{,}795$$

We prepare Table 12.6 for the analysis, entering these two sums of squares. By subtraction we find the ss *within* to be 12,634. In the last

TABLE 12.6

ANALYSIS OF VARIANCE OF GAINS IN MTAI SCORES—FOUR GROUPS OF SENIORS IN A TEACHERS' COLLEGE

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square
Total	49	16,429	——
Between treatments	3	3,795	1,265**
Within treatments	46	12,634	275

** In general use, double asterisk indicates significance at the 1 percent level. Likewise single asterisk is used to indicate significance at 5 percent level.

column of Table 12.6 we enter the mean squares (or variances). For the test of significance we compute $F = 1{,}265/275 = 4.60$. In the Table of Appendix I we find the 5 percent critical value of $F(3, 46)$ to be 2.81 and the 1 percent critical value to be 4.24. We would hence reject at either level the hypothesis that the true mean gains are equal.

The "sensitivity" of an experiment may be increased (that is, increase the chance of detecting small differences) by increasing the size of samples or by reducing the *within* mean square used in the denominator of the F ratio. The *within* mean square represents variance in X not accounted for by the design. It is often called *error variance* or *experimental error* since in this sense it represents variation regarding which the experiment provides no information. This variance may be reduced by improving the reliability of the measurement of X or by designing the experiment in such a manner that known sources of variability can be "controlled" and separated from it. One way of accomplishing the latter is to use more than one basis of classification of measures.

12.6 TWO BASES OF CLASSIFICATION WITH NO INTERACTION

The advantages of analysis of variance become particularly apparent when observations are classified on the basis of more than one factor. In a learning experiment, for instance, it may be advantageous to classify subjects on several factors which might contribute to the outcome variable, $X_{i\cdot}$. Subjects may be grouped on the basis of information on home background, general aptitude, previous learning experience, sex, teacher, and other factors which the experimenter wishes to "control" so that he will be able to examine "effects" of treatments in the knowledge that effects of these other factors are not mixed with them and so that he can reduce *experimental error*. On the other hand, the treatment itself may be a complex of two or more factors such as length of time for study, order of presentation, instructional materials used, time of day, time of semester, and the like, the effects of which the experimenter is interested in studying. Complex factorial designs are beyond the scope of this book. The theory is largely an extension of analysis of variance with only two bases of classification. Most of the remainder of this chapter is an introduction to experimental design covering only the essential features of analysis of variance in two-way classifications.

When observations are classified according to more than one variable, analysis of variance involves another type of component known as interaction. We will develop this concept by means of the simplest type of two-way classification, namely, the 2 × 2 table with equal numbers of cases in the cells.

The idea of interaction is similar to ideas in the analysis of a two-way classification of frequencies in Section 10.6. Chi square was used to test hypotheses of *independence* by comparing observed frequencies determined from the *row* and *column* totals.

Instead of frequencies, we are here dealing with a variable, the measure, X_{ijk}, for the ith individual in the jth row and the kth column. The procedure of determining what *means* to "expect" in each cell is similar to determining, as we did in Chapter 10, the *frequencies* to "expect" for each cell. In this case also we may use the marginal totals of the table.

Suppose, for instance, that in an experiment there were four groups of ten subjects each who had been assigned to two instructors and two treatments. Suppose furthermore that the instructor totals and means and the treatment totals and means, on a measure such as the Minnesota Teacher Attitude Inventory, were as follows:

Instructor	Treatment		Total	Mean
	Experimental	Control		
I	A	B	1,225	61.25
II	C	D	1,065	53.25
Total	1,181	1,109	2,290	
Mean	59.05	55.45		57.25

How would we estimate the means for each of the groups (cells, A, B, C, and D) on the basis of the above information? Without additional information we assume that the *treatment effect* is the same for both instructors and the *instructor effect* is the same for both treatments. We can use deviations of means from the general mean, 57.25. Cell A, for example, is in row I, which deviates 61.25 − 57.25 = +4.00 from the general mean. It is also in the first column which deviates 59.05 − 57.25 = +1.80 from the general mean. The group in cell A is hence in the high instructor category and in the high treatment category. The corresponding observed "effects" are +4.00 and +1.80. If the *row* effect is the same for both columns and the *column effect* the same for both rows, we would thus "expect" for cell A, a mean of 57.25 + 4.00 + 1.80 = 63.05. Similarly we estimate a mean for cell B to be 59.45, for cell C, 55.05, and for cell D, 51.45.

Letting $\bar{X}_j$ equal a row mean, $\bar{X}_k$, a column mean, and $\bar{X}$ the general mean, we may more simply compute the hypothetical means (expected under the assumption of equal row effect in columns and equal column effect in rows) as follows:

$$(\bar{X}_j - \bar{X}) + (\bar{X}_k - \bar{X}) + \bar{X} = (\bar{X}_j + \bar{X}_k - \bar{X}) \qquad (12.7)$$

For cell A, this would be 61.25 + 59.05 − 57.25 = 63.05.

If the observed means for the four cells are reasonably close to the above "expected" values, there is no "*interaction*" between the two factors, instructor and treatment. The presence of *interaction* would indicate different row effects in the two columns and different column effects in the two rows. We will now make the comparison. The actual forty observations appear in Table 12.7. From Table 12.7 we compute all observed

TABLE 12.7

DATA FROM EXPERIMENT IN TEACHER TRAINING

Instructor	Treatment Groups			
	Experimental		Control	
I	69	62	65	88
	56	85	69	69
	32	46	70	49
	48	42	75	78
	82	66	9	65
II	65	98	58	51
	29	74	15	54
	75	85	66	16
	54	2	60	62
	64	47	33	57

TABLE 12.8

TOTALS AND MEANS FOR DATA OF TABLE 12.7

Instructor	Treatment Group					
	Experimental		Control		Total	
	ΣX	$\bar{X}$	ΣX	$\bar{X}$	ΣX	$\bar{X}$
I	(A) 588	58.80	(B) 637	63.70	1,225	61.25
II	(C) 593	59.30	(D) 472	47.20	1,065	53.25
Total	1,181	59.05	1,109	55.45	2,290	57.25

means: row means, column means, and the four group means. These appear in Table 12.8. The cell means under the hypothesis of *no interaction* and the observed cell means appear together in Table 12.9.

TABLE 12.9

DEVIATIONS FOR INTERACTION EFFECT

Instructor	Item	Treatment Group	
		Experimental	Control
I	Observed mean	58.80	63.70
	Mean under hypothesis	*63.05*	*59.45*
	Deviation	−4.25	+4.25
II	Observed mean	59.30	47.20
	Mean under hypothesis	*55.05*	*51.45*
	Deviation	+4.25	−4.25

The difference in any group between the mean we would expect under the hypothesis of no interaction and the observed mean is the deviation representing *interaction effect*. It is seen to be 4.25 in Table 12.9. In a fourfold table the differences between observed means and the means under the no-interaction hypothesis are the same except for sign.

In order to test the hypothesis of *no interaction* it is necessary to rule out other possible explanations for the differential effects. The mathematical model which permits the test assumes randomness. If in assigning subjects to groups a bias was introduced so that the ten in group B were those most inclined to respond to the control treatment, we might thus have an explanation for the discrepancies. Presumably in the experiment this was avoided by the use of a table of random numbers or some similar device for assigning the forty subjects to the four groups and then randomly assigning the four groups to the cells A, B, C, D. This emphasizes the importance of care in planning experiments in order that tests of hypotheses can be validly made.

There remain two other possible explanations for the "apparent" interaction in Table 12.8, *real* effect and "sampling" effect. If we consider each of the four means as sample means from a whole population of such sample means, we see that the differential effects exhibited in the table fairly represent the true interaction in the *population* only if the sample means are close estimates of the population means. The apparent interaction may reflect a real interaction in the population or it may simply be due to sampling. We thus ask the same kind of question which we ask in any test of significance: Is it due to a real difference (real effect), or is it simply a difference due to sampling—a sampling effect? Squaring each one of the deviations, multiplying each by ten since there are ten subjects in each group, and summing over the four groups produce a sum of squares representing the observed departure from the no interaction hypothesis, *unless* these deviations may be explained by sampling. Since the magnitude of the deviation is the same for the four cells and since there is the same number of observations in each of the four cells, this computation simplifies to:

$$\text{ss } \textit{for interaction} = (4)(10)(4.25)^2 = 722.5$$

With the proper number of degrees of freedom we may compute from this a mean square to be used in testing the *no interaction* hypothesis. We will do this in the next section in which we will find that the interaction test fits nicely into the complete scheme of analysis for the data of Table 12.7. The analysis of this section is only for purposes of examining directly the sum of squares representing interaction. We will arrive at the same numerical result without the necessity of computing "expected" cell means, or for that matter any of the means, in the complete analysis of variance of the next section.

12.7 TWO BASES OF CLASSIFICATION— THE GENERAL CASE

Two distinct steps are required to make the complete analysis of the data in Table 12.7. The first step analyzes the data into the two components, (*a*) that *within* cells and (*b*) that *between* the four cells without regard to the form of the two-way classification. In this step we follow precisely the procedure of Section 12.3 in order to determine the total sum of squares and the sum of squares among the four means, subtracting to obtain the within sum of squares. Summing the squares of the individual scores in Table 12.7, we find that

$$\Sigma X_{ijk}^2 = 149{,}782$$

We compute the sum of squares for *total* as follows:

$$\text{ss } \textit{total} = 149{,}782 - (2{,}290)^2/40 = 18{,}680$$

The sum of squares between the four group means, computed without regard to the two-way classification, is as follows:

$$\begin{aligned}\text{ss between (four groups)} &= \tfrac{1}{10}\,[(588)^2 + (637)^2 + (593)^2 + (472)^2] - 131{,}102 \\ &= 1{,}493\end{aligned}$$

By subtraction we find the within group sum of squares to be 17,187. This and the total sum of squares we enter in Table 12.10, along with the corresponding degrees of freedom. There is a total of 40 subjects in the table. Therefore, the degrees of freedom for the total is 39. There are four groups in the analysis; they account for three degrees of freedom, which leaves 36 for the *within* sum of squares. Another way of computing the *within* degrees of freedom is to consider that for each of the four groups of ten subjects there are nine degrees of freedom and that if we pool the information from the four groups, we will have (4)(9) = 36 d.f.

TABLE 12.10

ANALYSIS OF VARIANCE FOR TABLE 12.7

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square
Total	39	18,680	—
Between treatments	1	130	130
Between instructors	1	640	640
Interaction	1	723	723
Within	36	17,187	477

We did not simply enter the sum of squares *between* in Table 12.10 because we can partition it into three components. The partitioning is the second step in the computations required for the complete analysis.

One component is that for *between treatments*. Referring to Table 12.8, we view the distribution of scores without regard to the instructor classification as though we had simply two groups of twenty each in the experimental and the control treatments, respectively. The computations are straightforward for finding the sum of squares between the means of these two groups as follows:

$$\text{ss } \textit{between treatments} = \tfrac{1}{20}\,[(1{,}181)^2 + (1{,}109)^2] - 131{,}102 = 130$$

Similarly, lumping together groups A and B for instructor one and groups C and D for instructor two, we may find the sum of squares *between instructors* as follows:

$$\text{ss } \textit{between instructors} = \tfrac{1}{20}\,[(1{,}225)^2 + (1{,}065)^2] - 131{,}102 = 640$$

These two sums of squares are for what are called "main effects"; we enter them in Table 12.10.

The total sum of squares *among all four means* we previously found to be 1,493. Subtracting from this the 130 which is accounted for by treatments and the 640 which is the sum of squares for instructors, the remainder, 723, is the sum of squares for interaction. Except for rounding, this is the same as the result of direct computation of the sum of squares for *interaction* in the previous section. Just as the sum of squares for *interaction* is the remainder of the sum of squares *between* (for all four groups) after subtracting the sum of squares for *treatments* and the sum of squares for *instructors*, so also the degrees of freedom for *interaction* may be found by subtraction. In this special case of the two-by-two design, the degrees of freedom among all four means is 3, and there is 1 d.f. each for treatments and instructors. This leaves 1 d.f. also for interaction.

Finally we compute the mean squares and enter them in Table 12.10. The *treatment* mean square is less than the *within* mean square so that the ratio would be less than one and could not indicate a treatment effect. Both the instructor effect and the interaction effect are greater than one, but neither is significant.

The mean square *within*, the *error variance*, is the common denominator of the F ratios computed to test various hypotheses. The smaller it is, the more sensitive the experiment will be in the sense that it will be more likely that any existing difference will be detected. The *within* or *error* variance is attributable to unknown causes, but by the nature of the design these causes are not treatment effect or instructor effect.

The number of subjects in two-way tables must be equal or proportional in order to test interaction conveniently. Otherwise it would not be possible to compute interaction effects directly as in Table 12.9. In such a table with disproportionate subclass numbers, the sums of squares for rows, columns, interaction, and within cells do not generally add to the total. There are methods of adjusting for lack of proportionality of frequencies in two-way classifications which permit tests of interaction.[1]

The two-by-two design with equal n is but a simple case of the $R \times C$ design with r rows and c columns and an equal number of individuals in each cell. The ideas of the simpler example carry over easily to the more general case. We now consider observations classified on the basis of one factor or characteristic into r rows (or categories), and on the basis of another factor or characteristic into c columns (or categories). Table 12.11 summarizes results of data from such a design. Thirty-two classrooms were chosen at random, one each from each of four grade levels in each of eight school systems. By means of a standardized classroom observation device, six different observations were obtained for

[1] See references 3 and 19 at the end of this chapter.

each of the thirty-two classrooms. One measure derived from the observations was called *differentiation*. This is roughly defined as the extent to which the classroom process accommodates the individual differences of pupils. Table 12.11 contains the totals of the thirty-two sets of six differentiation scores and the totals for rows (school systems) and the totals for columns (grades).

TABLE 12.11

Total of Six Differentiation Scores in each of Thirty-Two Classrooms, One Each in Four Grade Levels in Eight School Systems

School System	Grade				
	IV	VI	VIII	X	Total
1	96	74	63	66	299
2	62	68	57	84	271
3	86	94	68	60	308
4	66	64	69	66	265
5	75	74	63	39	251
6	61	75	67	46	249
7	66	72	64	66	268
8	77	71	66	50	264
Total	589	592	517	477	2,175

In the study from which these data were taken the objective was to study school system variation among a population of schools. The grade classification was intended as a "control." Note that grade level is not a random sample of grade levels. The grade levels were chosen simply to provide a spread over the upper elementary and lower secondary grades.

Before proceeding to the analysis of these data, let us review the theory of the procedure of partitioning the sums of squares. We let X_{ijk} represent the ith observation in the jth row and the kth column, $\bar{X}_{jk}$ the mean of the cell in the jth row and the kth column, $\bar{X}_j$ the mean of the jth row, $\bar{X}_k$ the mean of the kth column, $\bar{X}$ the general mean, n the number of observations per cell, r the number of rows, and c the number of columns.

As previously, the components of degrees of freedom and sums of squares (but not the mean squares) are additive. The first step in partitioning a sum of squares is based upon the same theory as equation 12.3 and is identical to it except for slightly more elaborate notation, as follows:

$$\sum_k \sum_j \sum_i (X_{ijk} - \bar{X})^2 = \sum_k \sum_j \sum_i (X_{ijk} - \bar{X}_{jk})^2 + n \sum_j \sum_k (\bar{X}_{jk} - \bar{X})^2 \qquad (12.8)$$

By a little algebra similar to that used in developing equation 12.3 it can be proved that the second member on the right side of equation 12.8 may be further subdivided:

$$n\sum_j\sum_k(\bar{X}_{jk} - \bar{X})^2 = n\sum_j\sum_k(\bar{X}_{jk} - \bar{X}_j - \bar{X}_k + \bar{X})^2$$
$$+ nc\sum_j(\bar{X}_j - \bar{X})^2 + nr\sum_k(\bar{X}_k - \bar{X})^2 \qquad (12.9)$$

The first component may be recognized from equation 12.7 to be the sum of squares for interaction. The other two are, respectively, sums of squares for rows and for columns.

There are $(N - 1) = (nrc - 1)$ degrees of freedom for total, $(r - 1)$ for rows, and $(c - 1)$ for columns. There are rc cells so that there are $(rc - 1)$ degrees of freedom for the sum of squares *between cell* means (without regard to classification). The interaction degrees of freedom is obtainable as a remainder thus:

$$\text{Interaction d.f.} = (rc - 1) - (\mathrm{r} - 1) - (c - 1) \qquad (12.10)$$
$$= (\text{d.f. cells}) - (\text{d.f. rows}) - (\text{d.f. columns})$$

Combining and factoring, we change this expression to

$$\text{Interaction d.f.} = (r - c)(c - 1) \qquad (12.11)$$
$$= (\text{d.f. rows})(\text{d.f. columns})$$

It is well as a check to use both equation 12.10, the interaction degrees of freedom as a remainder, and equation 12.11, the interaction degrees of freedom as the product of the degrees of freedom for the two main effects.

A diagram of computing instructions appears as Table 12.12. Note that row *b* of this table is the total of entries in rows *c*, *d*, and *e*. The sum of squares for rows *e* and *f* in this table may be computed from formulas easily derived from entries in the first four rows, although each may properly be computed by subtraction.

Two points should be especially noted about the formulas in Table 12.12 for computing sums of squares. In each case, the first term is the sum of squares of raw scores or totals of raw scores *divided by the corresponding number of observations*. For *total* the first term is a summation of the squares of individual raw scores. Consequently, the divisor is unity. For *between* cells the first term is a summation of squares of *cell* totals so the divisor is n, the number of observations per cell. For *rows* the divisor is observations per row; for columns, observations per column. The second point concerns the second term, which is seen to be common to the four formulas. This term, T^2/N, is sometimes called the *correction* since it "corrects" the sum of squares of raw scores to the sum of squares of "deviations."

TABLE 12.12

ANALYSIS OF VARIANCE WITH DOUBLE BASIS OF CLASSIFICATION AND EQUAL n

Source of Variation	Degrees of Freedom	Sum of Squares (Computing Form)
(*a*) Total	$N - 1$	$\sum_k \sum_j \sum_i X_{ijk}^2 - T^2/N$
(*b*) Between cells	$rc - 1$	$\sum_k \sum_j T_{jk}^2/n - T^2/N$
(*c*) Between rows	$r - 1$	$\sum_j T_j^2/nc - T^2/N$
(*d*) Between columns	$c - 1$	$\sum_k T_k^2/nr - T^2/N$
(*e*) $R \times C$ interaction	$(r - 1)(c - 1)$	(Subtract *c* and *d* from *b*)
(*f*) Within cells	$N - rc$	(Subtract *b* from *a*)

To save space the 192.individual measures are not shown in Table 12.11. In order to follow the computing scheme of Table 12.12, we need the sum of their squares over all cells, $\sum_k \sum_j \sum_i X_{ijk}^2$. This was reported to be 26,569. From Table 12.11 we see that the grand total, T, is 2,175. With this and other information at hand in Table 12.11, the sum of squares are computed as follows:

$$\text{ss } total = 26{,}569 - (2{,}175)^2/192 = 1{,}930.33$$

$$\text{ss } classrooms = \frac{(96)^2 + (74)^2 + (63)^2 + \cdots + (50)^2}{6} - (2{,}175)^2/192$$

$$= 152{,}175/6 - 24{,}638.67 = 723.83$$

$$\text{ss } systems = \frac{(299)^2 + (271)^2 + \cdots + (264)^2}{24} - 24{,}638.67 = 130.21$$

$$\text{ss } grades = \frac{(589)^2 + (592)^2 + (517)^2 + (477)^2}{48} - 24{,}638.67 = 198.89$$

The sum of squares for *within* and *interaction* are found by subtraction to be 1,206.50 and 394.73, respectively. These values and the degrees of freedom are entered in Table 12.13, and the corresponding mean squares are computed. The data for "between classrooms" are not recorded in the table since they are of interest only in computing.

TABLE 12.13

ANALYSIS OF VARIANCE OF DATA IN TABLE 12.11

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square
Total	191	1930.33	—
School systems	7	130.21	18.60
Grades	3	198.89	66.30
$S \times G$ interaction	21	394.73	18.80**
Within classrooms	160	1,206.50	7.54

** Double asterisk indicates significance at 1 percent level.

Although our primary interest may be in finding out whether or not there are significant variations in differentiation scores among school systems and among grade levels (the main effects), it is important to examine the interaction effect.

12.8 WHEN THE INTERACTION IS SIGNIFICANT

The F ratio for testing the interaction of Table 12.13 is $18.80/7.54 = 2.49$. This is in the .01 rejection region for 21 and 160 degrees of freedom, and we would declare the interaction significant at that level. This outcome appreciably alters the interpretation of the remaining mean squares and the procedure for testing their significance. A sketch of major points in the theory underlying the $R \times C$ analysis of variance will help to make this clear.

First we note, for each cell, the following four plausible deviations:

Row effect: $a_j = \mu_j - \mu$; characteristic of the class in the jth row; varies from row to row, but the same for all cells in the jth row.

Column effect: $b_k = \mu_k - \mu$; characteristic of the class in the kth column; varies from column to column, but the same for all cells in the kth column.

Interaction: $(ab)_{jk} = \mu_{jk} - \mu_j - \mu_k + \mu$; characteristic of a particular cell; varies from cell to cell, but the same for all observations in a cell.

Error: $e_{ijk} = X_{ijk} - \mu_{jk}$; "error of observation"; varies from observation to observation.

We next examine the model used in developing the theory for testing the three possible hypotheses in this type of design. In terms of the general population mean, μ, and the above deviations, the population mean of any cell is assumed to be:

$$\mu_{jk} = \mu + a_j + b_k + (ab)_{jk}$$

Then for any single observation

$$X_{ijk} = \mu + a_j + b_k + (ab)_{jk} + e_{ijk} \qquad (12.12)$$

Recapitulating, the three hypotheses are:

1. *That the row means are equal.* This is the same as the hypothesis that the row effects are zero, that is, $H : \mu_j = \mu$, or $H : a_j = 0$.

2. *That the column means are equal.* This is the same as the hypothesis that the column effects are zero, that is, $H : \mu_k = \mu$, or $H : b_k = 0$.

3. *That there is no interaction.* This is the hypothesis that for all cells the interaction is zero, that is, $H : (ab)_{jk} = 0$.

If the hypothesis of *no interaction* is true, the components for any single observation may be expressed simply as

$$X_{ijk} = \mu + a_j + b_k + e_{ijk} \qquad (12.13)$$

In this case the effects of columns and rows are simply additive. In case the interaction is *not* zero, the row effect in equation 12.12 is partly reflected in a_j and partly in the combined effect of rows and columns, $(ab)_{jk}$. Likewise the effect of columns is partly b_k and partly the mixture of row and column in $(ab)_{jk}$.

If all three hypotheses are true, equation 12.12 reduces to

$$X_{ijk} = \mu + e_{ijk}$$

and the variance of X_{ijk} from a sample is then just the variance of the e_{ijk} and is an unbiased estimate of σ^2, the variance of the population.

Now suppose that all three hypotheses are false. In this case there will be row effects, column effects, and interaction. If we may assume a whole population of row effects, column effects, and interactions, their variances will be σ_a^2, σ_b^2, and σ_{ab}^2, respectively.[1] As before, σ^2 is the variance of e_{ijk} and is assumed to be the same for all cells. Suppose, furthermore, that we compute a sample set of *mean squares*, using the scheme of Table 12.12 and dividing sums of squares by corresponding degrees of freedom as in Table 12.13. It may be demonstrated that the expected value of these squares are as shown in Table 12.14.

TABLE 12.14

POPULATION VALUES ESTIMATED BY MEAN SQUARES IN $R \times C$ ANALYSIS OF VARIANCE WITH EQUAL n IN CELLS—ALL NULL HYPOTHESES FALSE*

Source of Variation	Expected Value of Mean Square
Rows	$\sigma^2 + n\sigma_{ab}^2 + cn\sigma_a^2$
Columns	$\sigma^2 + n\sigma_{ab}^2 + rn\sigma_b^2$
Interaction	$\sigma^2 + n\sigma_{ab}^2$
Within (error)	σ^2

* See footnote below. The variances in this table represent population *mean squares* if categories are fixed.

In statistical language, a statistic is an unbiased estimate of a parameter if its mathematical expectation equals the value of the parameter. The sample mean squares are thus unbiased estimates of the expectations (population values) shown in Table 12.14. We note that the observed row mean square is an estimate of a sum of variances which include the error variance, the variance due to interaction, and a variance due to row

[1] The assumption must be carefully noted. It is an invalid assumption in many experiments. In the analysis of Table 12.11, the four grade levels were specified and may not be regarded as a random sample from a whole population of grade levels. The column effects are thus *fixed*. So also would the effects of "method" be fixed in a study comparing, for instance, three particular methods of teaching spelling. Inferences concerning them must be restricted to the particular methods used in the experiment. In such cases it is more appropriate to represent the variation of such *fixed* effects as $\Sigma(\mu_j - \mu)^2/(r - 1)$ should they correspond to rows, or $\Sigma(\mu_k - \mu)^2/(c - 1)$ should they correspond to columns; not σ_a^2 or σ_q^2. The latter symbols are used in this chapter for simplicity, with the above distinction in mind.

effect. An F ratio computed with the sample *row mean square* in the numerator and the sample *within mean square* in the denominator may thus be affected by a combination of two causes, real differences among rows, and interaction. The test would therefore not be a valid one with regard to row effects alone.

It is appropriate to use the F ratio only to compare two sample variances which by hypothesis are estimates of the same population variance, or of two population variances which are equal by hypothesis. Under the hypothesis of no row effect the third term of the expression for rows in Table 12.14 would be zero. However, if there is interaction, the numerator of the F ratio would be an estimate of $(\sigma^2 + n\sigma_{ab}^2)$. This exceeds the expected value of the denominator by the second term, the interaction component. On the other hand, if we were to use the interaction mean square in the denominator, we have a ratio of two estimates of the same thing. Hence the interaction mean square is used for testing row effects when the presence of the interaction is known or suspected. If it is known that an interaction *does not* exist, it is appropriate to omit testing it and use the *within mean square* as error for testing row effects.

When, in the absence of prior knowledge of it, the interaction is found *not* significantly different from the within mean square, the σ_{ab}^2 component is assumed to be zero and the interaction and within mean squares are both measures of the same thing, σ^2. In this case, either the *within mean square* is used by itself as "error" mean square or pooled with *interaction mean square*. The pooled *mean square* is computed as the ratio of the *sum of squares interaction* plus *sum of squares within* to the *interaction degrees of freedom* plus *within degrees of freedom*.

What has been said about testing row means applies equally to testing column means. For our purposes, the major point of Table 12.14 is that the *mean square* for either columns or rows estimates a value which may be greater than σ^2 for either or both of two reasons: (1) the effects of the corresponding factor are not zero, and (2) the interaction is not zero.

Since we found the interaction in Table 12.13 to be significant, the above analysis shows that we should use the interaction mean square as error mean square (denominator in the F ratio) in testing the hypotheses concerning the rows (school systems) and columns (grade level). The F for testing grades is $66.30/18.80 = 3.53$. For 3 and 21 d.f. this would be found to be significant at the 5 percent level but not at the 1 percent level. The F for testing school systems is obviously not significant. Note that by these tests we would accept the hypothesis of *no effect* for both rows and columns at the 1 percent level, and reject only in the case of grades at the 5 percent level.

It is of interest to compare these results with those using the within mean square for error in the F ratio. The systems ratio would be $18.60/7.54 = 2.47$. This is significant at the 5 percent level. The F for testing grades would be $66.30/7.54 = 8.79$. This is significant at the 1 percent level. Notice that using interaction (the larger) mean square for error we are not using an unbiased estimate of σ^2, but an unbiased estimate of the row mean square on the hypothesis of no row effect. Also notice that the significant interaction shows that both rows and columns have effects, and furthermore that the row effect fluctuates from column to column, and conversely. A more thorough treatment of the question of what mean square to use for error in analysis of variance may be found in references 1 and 9.

12.9 TWO-WAY CLASSIFICATION—SINGLE ENTRY IN EACH CELL

Of special interest is one variation of the $R \times C$ analysis of variance design, the two-way classification with a single entry in each cell. The theory of the last two sections is varied in only one important respect. With only one entry in a cell, it is not possible to compute a *within* mean square to use as error variance.

TABLE 12.15

DATA IN FOUR A CATEGORIES AND THREE B CATEGORIES—SINGLE ENTRY IN A CELL

A Category	B Category			T_j	$\bar{X}_j$
	(1)	(2)	(3)		
(1)	2	4	4	10	3.33
(2)	3	2	4	9	3.00
(3)	1	3	2	6	2.00
(4)	2	3	5	10	3.33
T_k	8	12	15	(35)	
$\bar{X}_k$	2.00	3.00	3.75		(2.92)

Since it has many applications in educational research, we will discuss this design with a simple hypothetical example to illustrate the computation methods and statistical tests. The data appear in Table 12.15 as twelve observations, one for each combination of four A categories and three B categories. A choice from a large number of possible real factors may be imagined for the two bases of classification, A and B. Here are a few possibilities:

A (rows)	B (columns)	Variable (X)
Four different aptitude levels.	Three methods of instruction.	Achievement scores for twelve different pupils.
Four different pupils	Three trials for each pupil.	Twelve measures of success or failure on task.
Four methods of presentation.	Three variations in subject-matter content.	Gain of twelve different subjects in knowledge.
Four different classes of school district.	Three levels of financial ability.	Library books per pupil in twelve school districts.
Four different bus drivers.	Three makes of school bus.	Unit operating cost over stated period.
Four different economic classes of family.	Three types of high-school program of study.	First-year college grade-point average of twelve graduates.
Four pupils.	Three test items.	Test item score.

In some of these examples one of the factors may be introduced for purposes of “control.” Selection and assignment of elements are presumably random. In some instances the treatment or other factor may be considered random variables from a population of effects; in other instances they may be *fixed* so that inferences would have to be restricted to the particular categories used in the experiment. In actual practice the design would be modified appropriately for these examples. For instance, we would need (and probably could easily get) more than four pupils and more than three items for the last example.

The variation of the twelve measures may be due to row effect (A), column effect (B), and, *assuming no interaction*, experimental error. The last is measured by the population variance independent of row and column effects. Dropping the i subscript in equation 12.12, since there is only one observation per cell, the basic model is

$$X_{jk} = \mu + a_j + b_k + e_{jk} \qquad (12.14)$$

Modifying Table 12.14 by assuming $\sigma_{ab}^2 = 0$ and putting $n = 1$, the expectations of the three mean squares which may be computed for this design are

$$\begin{array}{ll} \text{Rows:} & \sigma^2 + c\sigma_a^2 \\ \text{Columns:} & \sigma^2 + r\sigma_b^2 \\ \text{Error:} & \sigma^2 \end{array} \qquad (12.15)$$

The third line is now named *error* (or sometimes *residual*) since it is an unbiased estimate of σ^2. The fourth line of Table 12.14 is omitted since the data permit no estimate of a "within" cell mean square. Under the hypothesis of no interaction we use the error (or residual) mean square for testing row and column effects.

Instead of a cell mean, $\bar{X}_{jk}$, based on n observations, as in equations 12.8 and 12.9, there is but one observation, X_{jk}, in the jth row and the kth column. The first member on the right of equation 12.8 is zero, leaving only the second, $\sum_j \sum_k (X_{jk} - \bar{X})^2$. This is the same as the sum of squares for the total. By similar changes, equation 12.9 becomes

$$\sum_j \sum_k (X_{jk} - \bar{X})^2 = \sum_j \sum_k (X_{jk} - \bar{X}_j - \bar{X}_k + X)^2 + c\sum_j (\bar{X}_j - \bar{X})^2 + r\sum_k (\bar{X}_k - \bar{X})^2 \qquad (12.16)$$

The first term on the right is the error or residual sum of squares and may be computed as a remainder by subtracting *sum of squares rows* and *sum of squares columns* from *sum of squares total*. The symbols show it to be (like interaction in Sections 12.6 and 12.7) the sum of squares of deviations of observed cell values after row and column effects are taken into account. Equation 12.14 indicates how this fits into the analysis of variance model of additive components in the population and that the error is to be computed as

$$e_{jk} = X_{jk} - (\mu + a_j + b_k)$$

Substituting sample estimates for the three terms in the parentheses, we can estimate e_{jk} by

$$\hat{e}_{jk} = X_{jk} - [\bar{X} + (\bar{X}_j - \bar{X}) + (\bar{X}_k - \bar{X})] = (X_{jk} - \bar{X}_j - \bar{X}_k + \bar{X}) \quad (12.17)$$

The caret indicates that the expression is an estimate of the error based upon sample information. The sum of squares of the estimated errors is the first term on the right of equation 12.16.

TABLE 12.16

ANALYSIS OF VARIANCE OF TABLE 12.15

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	F
Total	11	14.92	—	—
A (rows)	3	3.59	1.20	1.40
B (columns)	2	6.17	3.08	3.58
Error (residual)	6	5.16	.86	—

The analysis of variance appears in Table 12.16, two decimals being retained to illustrate arithmetic details. The best computation scheme is implied in equation 12.16, the error sum of squares being found by subtraction. Since $N = rc$, and there are $(r - 1)$ degrees of freedom for rows and $(c - 1)$ degrees of freedom for columns, the error sum of squares has $(r - 1)(c - 1)$ degrees of freedom. The formal analysis appears in Table 12.17. Since neither the F for rows nor the F for columns is significant at the .05 level, we accept both null hypotheses concerning A and B effects.

TABLE 12.17

ANALYSIS OF VARIANCE WITH DOUBLE CLASSIFICATION AND SINGLE ENTRY IN EACH CELL

Source of Variation	Degrees of Freedom	Sum of Squares (Computing Form)
(a) Total	$N - 1$	$\sum_j \sum_k X_{jk}^2 - T^2/N$
(b) Between rows	$r - 1$	$\sum_j T_j^2/c - T^2/N$
(c) Between columns	$c - 1$	$\sum_k T_k^2/r - T^2/N$
(d) Residual	$(r - 1)(c - 1)$	(Subtract b and c from a)

Though the error sum of squares is most easily computed by subtraction, it can be found directly, and this sometimes provides a valuable check on arithmetic. Table 12.18 shows how the twelve measures of Table 12.15 would look if there was no error and each measure consisted only of the general mean, $\bar{X}$, plus A effect, $(\bar{X}_j - \bar{X})$, plus B effect, $(\bar{X}_k - X)$. These are the best estimates of cell means, μ_{jk}, afforded by the data. For example, the entry in row 1 and column 1 in Table 12.18 may be computed from Table 12.15, as follows:

TABLE 12.18

ESTIMATES OF CELL MEANS, μ_{jk}*

A Category	B Category			T_j	$\bar{X}_j$
	(1)	(2)	(3)		
(1)	2.41	3.41	4.16	9.98	3.33
(2)	2.08	3.08	3.83	8.99	3.00
(3)	1.08	2.08	2.83	5.99	2.00
(4)	2.41	3.41	4.16	9.98	3.33
T_k	7.98	11.98	14.98	34.94	
$\bar{X}_k$	2.00	3.00	3.75		2.92

* Each cell entry is its row deviation plus its column deviation plus the general mean, that is, $(\bar{X}_j - \bar{X}) + (\bar{X}_k - \bar{X}) + \bar{X} = (\bar{X}_j + \bar{X}_k - \bar{X})$.

$$2.92 + (3.33 - 2.92) + (2.00 - 2.92) = 2.41$$

The marginal totals and means of Tables 12.15 and 12.18 are the same except for rounding. Subtracting the entry in each cell of Table 12.18 from the corresponding observed score in Table 12.15 yields the deviations given in Table 12.19. The sum of these deviations in each row and each

TABLE 12.19

RESIDUALS FOR ERROR VARIANCE FROM TABLES 12.15 AND 12.18

A Category	B Category		
	(1)	(2)	(3)
(1)	−.41	+.59	−.16
(2)	+.92	−1.08	+.17
(3)	−.08	+.92	−.83
(4)	−.41	−.41	+.84

column is zero, except for rounding, and their sum of squares is 5.17, agreeing with the error sum of squares in Table 12.16, except for rounding errors.

Recalling the assumptions we made in the analysis of Table 12.16, we can understand the limitations of this design and recognize situations in which it is inappropriate. We assumed no interaction, so that the variance of the deviations shown in Table 12.19 is an unbiased estimate of the σ^2. This may be viewed as the variance we would find if all subjects were of the same row category and the same column category so that no row or column effect contributed to the total sum of squares. If there is interaction, this mean square is not an unbiased estimate of σ^2. Referring to Table 12.14, we see that its expected value is $\sigma^2 + \sigma_{ab}^2$, so that it is an unbiased estimate of the row mean square on the hypothesis of no row effect, and likewise for columns. Lacking data which provide an estimate of the within cells mean square, we cannot test interaction, and there is no way of knowing whether row and column effects are being tested against error or interaction.

It is preferable in designing exploratory studies in areas involving factors the relationships of which are unknown to use a design which involves more than one entry in each cell, as in the design of Table 12.11, so that interaction can be tested. If we add one more randomly assigned case per cell, $n = 2$, the result is the same as a complete repetition or "replication" of the experiment. If there are n entries in a cell, the experiment is said to be "replicated" n times.

12.10 TEST RELIABILITY AND ANALYSIS OF VARIANCE

Analysis of variance has important contributions to the theory of test reliability. If test scores are assumed to consist additively of several uncorrelated components, so that an individual score is the sum of these components, the total variance of the test can be partitioned into component variances. If it is assumed that one of these components is *error* and that another is the *true score*, the total variance of the test consists of the *true variance plus error variance*. By definition of the coefficient of reliability (equation 9.23), these components determine the reliability of a test. If we consider the true score of an individual to be the population mean, μ_i, of a hypothetical infinite number of administrations of the test, an individual test score, X_i, will then consist of the two components

$$X_i = \mu_i + e \qquad (12.18)$$

Assuming that the two components of X_i are independent, in the population of individuals,

$$\sigma_x^2 = \sigma_\mu^2 + \sigma_e^2 \tag{12.19}$$

We may express the coefficient of reliability in terms of these components as follows:

$$\rho_{tt} = \frac{\sigma_\mu^2}{\sigma_\mu^2 + \sigma_e^2} = 1 - \frac{\sigma_e^2}{\sigma_x^2} \tag{12.20}$$

If we have a sample of only two administrations of the same test (or scores from two separate but comparable forms), we would compute the coefficient of reliability by the correlation methods of Section 9.11. Imagine, however, that we have $n > 2$ test scores of some one characteristic from a random sample of individuals. These data could be arranged in a simple analysis of variance. This would permit us to estimate the variance *among* the individual mean scores, and the variance *within* individuals. These two variances are by definition estimates of (1) n times the variance of the *true* scores plus the error variance, and (2) the *error* variance respectively. From our sample set of observations we may use the *among individuals* mean square, ms_b, as an estimate of $\sigma_e^2 + n\sigma_\mu^2$ and the mean square within, ms_w, as the estimate of the error variance, σ_e^2.

By simple algebra it can be shown that the estimate of the coefficient of reliability, from equation 12.20 is

$$r_{tt} = \frac{ms_b - ms_w}{ms_b + (n - 1)ms_w} \tag{12.21}$$

This formula is of value in determining the reliability of ratings of individuals, products of individuals, or of objects. Suppose that there are four ratings for each of five classrooms, A, B, C, D, and E, on some characteristic as follows: (A) 4, 5, 4, 6; (B) 7, 6, 7, 6; (C) 8, 9, 7, 9; (D) 10, 9, 10, 9; and (E) 12, 11, 11, 10. These five sets of ratings appear consistent in the sense that the variation *within* is low relative to the variation *between*. Another way to put it is that there appears to be a correlation (correspondence) of ratings. Equation 12.21 in fact will measure this correlation. It is the average of all possible correlation coefficients, r, computed by taking five pairs at a time, one pair each from each of the five groups. As a measure of the correlation of individuals *within* groups, it is called *intraclass* correlation.

The sum of squares of the 20 classroom ratings is 106.0, the sum of squares *between* is 96.5, and the sum of squares *within* is 9.5, the degrees of freedom between is 4, and within 15. Thus, $ms_b = 24.12$ and $ms_w = 0.63$. Substituting in equation 12.21, we have

$$\begin{aligned} r_{tt} &= (24.12 - 0.63)/(24.12 + 1.89) \\ &= (23.49)/(26.01) = .90 \end{aligned}$$

Should interest lie not in the reliability of the *individual rating*, but in the reliability of the *average* of the four ratings, the above may be adjusted by the Spearman-Brown formula (9.29), with $k = 4$, or, by substituting equation 12.21 for ρ_{12} in the Spearman-Brown formula, we may easily derive the following formula for finding the reliability of the mean rating when there are two or more ratings or measures for each subject:

$$r_{nn} = \frac{ms_b - ms_w}{ms_w} \tag{12.21a}$$

The reliability of the above classroom *mean ratings* is estimated by this to be

$$r_{nn} = (24.12 - 0.63)/(24.12) = .97$$

Of special interest in measurement theory is the $R \times C$ design of the previous section applied to repeated trials or different measures at different times or several measures taken at the same time. If the measures are spaced over time, differences between them reflect absence of *stability* in the function measured or in its measurement. If the measures are taken at the same time, differences between them reflect absence of *equivalence* reliability. If rows represent individuals and columns represent different but parallel measures of the same individuals, we have a two-way classification which would permit a different method of estimating the coefficient of reliability and the error of measurement. It is a much-used formula for the *equivalence* reliability of a test. It is also of value as a measure of reliability of ratings (as in the foregoing example) which takes "between rater" variance into account if ratings are by different individuals. We will develop the formula from two-way analysis of variance. First we assume that the columns represent items so that the cell entries contain item scores (or subscores) of a total test. We assume furthermore that the total scores of individuals on the test (represented by the row totals) consist of three independent components of variation: (1) that associated with items (column effect); (2) that associated with individuals (row effect); and (3) an error component not correlated with the other two. Noting that there is no replication, and making no assumption about interaction in this design, we find from Table 12.14 the following expected values of the mean square among individuals (rows) and the residual mean square:

$$E(ms_a) = \sigma_e^2 + \sigma_{ab}^2 + c\sigma_a^2 \tag{12.22}$$

$$E(ms_r) = \sigma_e^2 + \sigma_{ab}^2 \tag{12.23}$$

In analysis of variance, sums of squares and mean squares are usually expressed in terms of units of the individual data. In the design we are discussing the mean squares are all in terms of the X_{ijk}, which are individual measures of a person on one of the parts (items or subtotals) of the test. As was shown in the first three sections of this chapter, the comparison of such mean squares tests whether there are differences among means. It is the total test score which is of interest to the test analyst, and differences among totals are tested by precisely the same test.[1] We may therefore use the analysis of variance approach to test the hypothesis of no difference in total scores among individuals. This would tell us whether or not to expect differences in true scores among individuals in the population, but not *how well* the measure discriminates among them. However, a reliability coefficient may be computed from the variance components. The F test of individuals is a test of the significance of this reliability coefficient.

If we consider the expected value of ms_r as *error variance*, and the expected value of ms_a as the variance *among* individuals, their difference will represent the variance of *true* scores. By definition, the coefficient of reliability is then

$$\rho_{tt} = \frac{c\sigma_a^2}{\sigma_e^2 + \sigma_{ab}^2 + c\sigma_a^2} \tag{12.24}$$

and, substituting sample statistics, it is estimated by

$$r_{tt} = \frac{ms_a - ms_r}{ms_a} \tag{12.25}$$

Note that by defining mean squares in both instances as mean square *between* and mean square for *error*, equations 12.21*a* and 12.25 are the same. The procedure differs only in the method of determining the mean square for *error*.

An example of the application of equation 12.25 is drawn from a study of the relationship of pupil-teacher attitudes and satisfaction with student teaching. In that study a satisfaction scale consisting of 32 items was administered to 177 women elementary student teachers.[2] The layout,

[1] We could easily convert all sums of squares and all mean squares to the "total" dimension by multiplying each score by c, the number of parts (or items) in the total. Any row mean would then be $\Sigma cX_{ijk}/c = \Sigma X_{ijk}$ = total score. All mean squares would be c^2 times as large as normally. All F ratios would however be unchanged, since c^2, the factor common to all mean squares, would cancel. Simple algebra will prove this and the general principle that all basic elements in an analysis of variance may be multiplied or divided by a constant, or increased (or decreased) by a constant without altering the significance tests based upon ratios of mean squares.

[2] Harold E. Mitzel and Louis P. Aikman, *Teacher-Pupil Attitude and Their Relationship to Satisfaction with Student Teaching*, New York, College of the City of New York, Publication 20, Division of Teacher Education, Office of Research and Evaluation (mimeographed), May 1954.

not reproduced here, may be viewed as consisting of 177 rows (individuals), and 32 columns (items), with a single item score in each cell. An analysis of variance of the results appears as Table 12.20. Using equation 12.24, we estimate ρ_{tt} from equation 12.25 as follows:

$$r_{tt} = 1.00 - (.140)/(.900) = .84$$

TABLE 12.20

ANALYSIS OF VARIANCE OF SATISFACTION SCALE SCORES*

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square
Total	5,663	966.28	—
Individuals	176	158.34	.900
Items	31	45.79	1.477
Error (residual)	5,456	762.15	.140

* From Mitzel and Aikman, footnote 2, p. 284

An essentially equivalent form of equation 12.25 for the coefficient of reliability is the formula known as Kuder-Richardson formula 20,

$$r_{tt} = \frac{c}{c-1}\left(\frac{s_t^2 - \Sigma p_k q_k}{s_t^2}\right) \tag{12.26}$$

$$= \frac{c}{c-1}\left(1 - \frac{\Sigma p_k q_k}{s_t^2}\right)$$

where $c =$ number of items; $s_t^2 =$ variance of the T_j, or total scores of individuals; $p_k =$ proportion passing item k; and $q_k = (1 - p_k)$.

Some of the features of equation 12.26 can be explored by assuming that a large number of individuals have taken the test so that the difference between r and $(r - 1)$ is negligible, that every individual attempts every item, and that an item score is either 1 or 0. A total score on the test is the sum of item scores, the number of correct responses. Our design is now one of r individuals (rows) and c items (columns), with scores of 1 and 0 in cells. In this situation

$$\bar{X}_k = T_k/r = p_k$$

where T_k is the number of subjects passing the kth item and p_k is the proportion of subjects passing the kth item. Since item scores in the cells must be either 1 or 0, $X_{jk}^2 = X_{jk}$. Then if we compute the variance of an item by dividing by r (disregarding the difference between r and the proper degrees of freedom, $r - 1$), we find that

$$\begin{aligned} s_k^2 &= \frac{\sum\limits_j X_{jk}^2}{r} - \left(\frac{\sum\limits_j X_{jk}}{r}\right)^2 \\ &= \frac{T_k}{r} - \left(\frac{T_k}{r}\right)^2 \\ &= p_k - p_k^2 = p_k(1 - p_k) \\ &= p_k q_k \end{aligned}$$

Now suppose that we compute the correlation of each item in turn with every other item. With scores of 1 and 0, we have two-point distributions, and the correlation coefficient is a phi coefficient (Section 13.18). Then for each pair of items there is a covariance, $s_g s_k r_{gk}$. It can be shown that the total variance consists of the sum of the item variances, $\Sigma p_k q_k$, and the covariances. Formula 12.26 for computing reliability contains the sum of the item variances as error variance. The remainder of the total is the sum of the inter-item covariances. This sum represents the *true* variance among individuals.

Since this reliability formula is a constant multiplied by the ratio of the sum of item covariances to the total variance among individuals, it appears that the reliability depends upon the item intercorrelations. The higher the intercorrelations, the greater r_{tt}. Hence it is a measure of the *homogeneity* of items in the test, that is, a measure of *internal consistency* of measurement at a particular point of time.

A reliability coefficient computed from the Kuder-Richardson formula 20 (equation 12.26) may be considered as the expected value of reliability coefficients from random halves of items (the average of all possible combinations of halves). The derivation is based upon the assumption of two parallel tests with equal variances and equal average item covariances, as well as equality of average item covariances within tests and equality of average covariances of items of one test with items of another.

More elaborate designs of analysis of variance have been used in determining the reliability of measurement. Reference 15 at the end of this chapter may be consulted for more advanced applications of analysis of variance to testing.

12.11 ANALYSIS OF VARIANCE WITH ONLY TWO GROUPS

The t test of Chapter 11 for testing the significance of differences between two means is but a special case of the more general methods of analysis of variance. With a single criterion of classification, into only two groups, it is easily demonstrated that the F ratio of analysis of variance is the square of the t as used in equation 11.15, so that the two tests are identical.

If the two sample sizes are n_1 and n_2, the corresponding totals T_1 and T_2, and the corresponding sample means $\bar{X}_1$ and $\bar{X}_2$, we may write in abbreviated notation

$$\Sigma x_1^2 = \sum_{i=1}^{n_1} (X_{i1} - \bar{X}_1)^2$$

and

$$\Sigma x_2^2 = \sum_{i=1}^{n_2} (X_{i2} - \bar{X}_2)^2$$

The sum of squares *within* is

$$ss_w = \Sigma x_1^2 + \Sigma x_2^2$$

and the degrees of freedom *within* are $(n_1 + n_2 - 2)$. Hence,

$$ms_w = \frac{\Sigma x_1^2 + \Sigma x_2^2}{n_1 + n_2 - 2} \tag{12.27}$$

There is but 1 d.f. for the sum of squares between groups so that

$$ss_b = ms_b$$

The sum of squares *between* may be computed from

$$\begin{aligned} ss_b &= \frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} - \frac{(T_1 + T_2)^2}{n_1 + n_2} \\ &= \frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} - \frac{(T_1^2 + T_2^2 - 2T_1T_2)}{n_1 + n_2} \end{aligned}$$

With $n_1n_2(n_1 + n_2)$ as the lowest common denominator, we combine fractions and collect terms to obtain

$$ss_b = \frac{n_2^2T_1 + n_1^2T_2^2 - 2n_1n_2T_1T_2}{n_1n_2(n_1 + n_2)}$$

Rearranging,

$$\begin{aligned} ss_b &= \frac{n_1n_2}{(n_1 + n_2)}\left(\frac{T_1^2}{n_1^2} + \frac{T_2^2}{n_2^2} - \frac{2T_1T_2}{n_1n_2}\right) \\ &= \frac{n_1n_2}{(n_1 + n_2)}\left(\frac{T_1}{n_1} - \frac{T_2}{n_2}\right)^2 \end{aligned}$$

Hence,

$$ss_b = \frac{n_1n_2}{(n_1 + n_2)}(\bar{X}_1 - \bar{X}_2)^2 \tag{12.28}$$

From equations 12.27 and 12.28 we find that we may express the F ratio for testing means in the following form:

$$F = \frac{ms_b}{ms_w} = \frac{\dfrac{n_1 n_2}{(n_1 + n_2)}(\bar{X}_1 - \bar{X}_2)^2}{\dfrac{\Sigma x_1^2 + \Sigma x_2^2}{n_1 + n_2 - 2}}$$

$$= \frac{(\bar{X}_1 - \bar{X}_2)^2}{\left(\dfrac{\Sigma x_1^2 + \Sigma x_2^2}{n_1 + n_2 - 2}\right)\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)} \qquad (12.29)$$

From equations 11.14 and 11.15 it is evident that equation 12.29 is the square of t. Recalling that the appropriate t test is a two-tailed test, we may enter the t table of Appendix E for various degrees of freedom and square two-tailed 5 percent and 1 percent critical values of t values. These will be the same as values of F for the same number of degrees of freedom appearing in the first column (for 1 d.f. numerator) of the F table of Appendix I. Conversely, the square root of an F value in the first column of Appendix I is the two-tailed 5 percent and 1 percent critical values of the t table for the same number of degrees of freedom. For example, the 5 percent point in Appendix I for $F(1, 10)$ is 4.96. The square root of this is 2.23, which is the entry for $t_{.975}$ with 10 d.f.

TABLE 12.21

ANALYSIS OF VARIANCE FOR EXAMPLE OF SECTION 11.3

Source of Variation	Degrees of Freedom	Sum of Square	Mean Square
Total	21	3,337	—
Between	1	477	477
Within	20	2,860	143

Reworking the example of Section 11.3 concerning the CTMM scores of boys in high schools B and C from the material in Appendix G, we make a numerical check on theory. In Section 11.3 we found t to be -1.82. This was not sufficient to reject the hypothesis of no difference at the

5 percent level. The analysis of variance of the same data appears in Table 12.21. The F for testing means is $477/143 = 3.34$. This falls short of the 4.35 critical value for $F(1, 20)$ at the 5 percent level. The results of the two methods agree except for rounding, the square of the previously found value of t being 3.31.

The $R \times C$ design of Section 12.9 may be viewed as an extension of the "matched pairs" experiment of Chapter 11. One method of conducting an experiment is to select c individuals at random from each of r categories of a *control variable*. In a learning experiment, for instance, these categories might represent intelligence levels, levels of previous learning experience, amount of previous training, or some other classification known or assumed to be related to the criterion measure, X. If the c subjects in the r control classes (rows) are randomly assigned to the c experimental treatments, there will be c sets of subjects "matched" on the control classification. This is an $R \times C$ design with a single entry in each cell.

The example of Section 11.2 is but a special case, with $c = 2$. The data of Table 11.2 may be reworked by analysis of variance as in Table 12.22. The F ratio for testing treatments is $41.47/2.26 = 18.3$. The square root of this is 4.28, agreeing with the value of t found by means of equation 11.7 in Section 11.2. Hence we see that in comparing the two groups when *matched* in an $R \times 2$ design (or a $2 \times C$ design), with 1 d.f. for the numerator of the F ratio, $F = t^2$, and the tests are identical.

TABLE 12.22

ANALYSIS OF VARIANCE OF DATA IN TABLE 11.2

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square
Total	19	188.17	—
Subjects	9	126.32	14.04
Treatments	1	41.47	41.47
Error (residual)	9	20.38	2.26

The proof of this is of interest since it adds meaning to the theory of Chapter 11. In this case the F ratio has the mean square between columns in the numerator and the residual mean square in the denominator. The between columns sum of squares may be computed as in equation 12.28, but since there are the same number of observations, r, in each group, equation 12.28 becomes

$$ss_b = \frac{r^2}{2r}(\bar{X}_1 - \bar{X}_2)^2$$

$$= \frac{r}{2}\bar{D}^2 \tag{12.30}$$

The residual sum of squares is

$$ss_r = \sum_j (X_{j1} - \bar{X}_j - \bar{X}_1 + \bar{X})^2 + \sum_j (X_{j2} - \bar{X}_j - \bar{X}_2 + \bar{X})^2$$

where X_{j1} is the entry in the first column in the jth row, and X_{j2} is the entry in the second column of the jth row. Substituting $(X_{j1} + X_{j2})/2$ for $\bar{X}_j$ and substituting $(\bar{X}_1 + \bar{X}_2)/2$ for $\bar{X}$, squaring, and collecting terms, this reduces to

$$ss_r = \frac{1}{2}\sum_j [(X_{j1} - X_{j2}) - (\bar{X}_1 - \bar{X}_2)]^2$$

In the notation of Sections 11.1 and 11.2,

$$ss_r = \frac{1}{2}\sum_j d^2 \tag{12.31}$$

The number of degrees of freedom for *between* columns is 1, and the degree of freedom for *residual* is $(r - 1)$. From equation 12.30 the mean square for columns is

$$ms_b = \frac{r}{2}\bar{D}^2$$

From equation 12.31 the mean square for residual is

$$ms_r = \frac{\frac{1}{2}\Sigma d^2}{(r-1)}$$

Therefore, the F ratio for testing columns is

$$F = \frac{(r/2)\bar{D}^2}{\dfrac{\frac{1}{2}\Sigma d^2}{(r-1)}} = \frac{\bar{D}^2}{\dfrac{\Sigma d^2}{r(r-1)}}$$

From equations 11.6 and 11.7,

$$F = \frac{\bar{D}^2}{s_{\bar{D}}^2} = t^2 \tag{12.32}$$

12.12 ASSUMPTIONS UNDERLYING ANALYSIS OF VARIANCE

There has been frequent reference in this chapter to assumptions underlying the use of the F test in analysis of variance. Since a failure of these assumptions may vitiate the analysis, the circumstances under which the tests of this chapter are valid should be kept in mind in the earliest stages of planning an inquiry. In summary, the underlying assumptions are as follows:

(1) Individuals or observations in the groups are random samples under the null hypothesis.

(2) In designs with more than one basis of classification the effects are additive.

(3) The experimental errors are independently distributed.

(4) The experimental errors are normally distributed.

(5) There is homogeneity of variance of experimental errors among subgroups.

These assumptions are not unrelated. Failures usually occur in combination. In general, failure to satisfy these assumptions will introduce an error into the probability of rejecting a hypothesis, though such an error is unlikely to be large unless the failure is gross. Hence, when it is suspected that the assumptions are not satisfied, the significance levels are inexact. This problem was encountered in Section 11.6, in comparing means of two groups. The discussion in that section applies here to the extent that the two-group comparison is a special case of analysis of variance.

A few comments are appropriate as to how to avoid difficulties arising from failure of these assumptions. It is not always convenient or possible in educational research to plan experiments which permit satisfactory *randomization*. If the experimenter is limited to the treatment of a single intact class for treatment A, and another intact class for treatment B, he obviously cannot assign individuals randomly to treatments. The difficulty might be solved with adequate *replication* of his experiment by including additional classrooms. This, too, is not always possible. Nevertheless, assignment of individuals (or of whole classes) to categories is essential if the test of significance is to have a sound mathematical basis. Without randomization, no meaningful significance level can be calculated.

Randomization also assists in avoiding violations of the assumptions of independence, but it is not necessary for this purpose, and lack of independence of error from observation to observation may sometimes be avoided by an elaboration of the design so as to take into account such factors as time of observation and observer. Very skewed distributions may be accompanied by a failure of the effects to be additive.

Sampling experiments happily have shown that there may be considerable departure from the assumptions of normality and homogeneity of variance without seriously invalidating the F test. In a simple one-way design a visual examination of distributions of data within groups is usually an adequate precaution against this hazard. Unless the distribution is extremely unsymmetrical there will be no appreciable effect on the F test. As noted in Section 11.6, nonnormal distributions can sometimes be rectified by transformations such as the logarithm of X, the square root of X, or the inverse sine of X.

A visual comparison of distributions within groups may usually be sufficient for detecting serious heterogeneity of variance, but statistical tests are available if needed. They do not eliminate heterogeneity. They merely test the hypothesis that the variances of the parent subpopulations are the same. A short-cut test by Hartley involves computation of only the ratio of the highest variance to the lowest variance from among the variances of k groups. The loss in power compared to the more complex Bartlett test described below is small or negligible. The special tables needed for the Hartley test are to be found in references 10 and 20.

A Cochran test, also requiring special tables, uses a statistic which is simply the ratio of the largest variance to the sum of all the variances of the k groups. References 4 and 6 give the necessary tables.

The Bartlett test requires more computation. A statistic is found which is approximately distributed as χ^2, so that a χ^2 table may be entered to test the hypothesis

$$H : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \cdots = \sigma_k^2$$

This statistic is given by

$$M = \frac{2.3026}{C} [\nu \log s^2 - \Sigma(\nu_j \log s_j^2)] \tag{12.33}$$

where

$$\nu_j = (n_j - 1), \text{ the degree of freedom in the } j\text{th group;}$$

$$\nu = \Sigma\nu_j = N - k;$$

$$s_j^2 = (\Sigma x^2)/(n_j - 1);$$

$$s^2 = (\sum_j \nu_j s_j^2)/\nu = (\sum_j \sum_i x_i^2)/\nu; \text{ and}$$

$$C = 1 + \frac{1}{3(k-1)} \left[\sum \frac{1}{\nu_j} - \frac{1}{\nu}\right]$$

The factor, C, is a correction factor always larger than 1.00, and needs to be computed only if M' (M computed with $C = 1.00$) is found to be significantly large. M is approximately distributed as χ^2 with $(k - 1)$ degrees of freedom.

The computations for equation 12.33 are straightforward, and require only a table of logarithms and knowledge of their use. We will go through the steps of the computation, using the data of Table 12.5 to test the hypothesis of no difference in the variances of the four populations.

Step 1. Lay out a table showing for each group the degrees of freedom, the reciprocal of degrees of freedom, the sum of squares ($\nu_j s_j^2$), the variance, the logarithm of the variance, and the product of the degrees of freedom and the variance. Such a table for the data of Section 12.5 appears as Table 12.23.

TABLE 12.23

DATA FOR BARTLETT TEST OF HOMOGENEITY OF VARIANCES IN TABLE 12.5

Group	n_j	ν_j	$\dfrac{1}{\nu_j}$	Σx^2	s_j^2	$\log s_j^2$	$\nu \log s_j^2$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
I	9	8	.1250	780.22	97.53	1.9891	15.9128
II	18	17	.0588	5,750.28	338.25	2.5292	42.9964
III	11	10	.1000	4,114.55	411.46	2.6143	26.1430
IV	12	11	.0909	1,988.92	180.81	2.2572	24.8292
Total	50	46	.3747	12,633.97			109.8814

Step 2. Compute $\log s^2$. (In Table 12.23, divide the column 5 total by the column 3 total and find the logarithm of the result.)

$$\log s^2 = \log (12{,}633.97)/(46)$$
$$= \log 274.65 = 2.4388$$

Step 3. Find $\nu \log s^2$. (Multiply the result of step 2 by the total of column 3.)

$$\nu \log s^2 = (46)(2.4388) = 112.18$$

Step 4. Subtract $\Sigma(\nu_j \log s_j^2)$ from the result of step 3. (In Table 12.23, subtract the total of column 8 from the result of step 3.)

$$\nu \log s^2 - \Sigma(\nu_j \log s_j^2) = 112.18 - 109.88 = 2.30$$

Step 5. Multiply step 4 results by 2.3026,

$$M' = (2.3026)(2.30) = 5.30$$

Step 6. Enter a chi-square table (with d.f. $= k - 1$) at the critical value for a predetermined level of risk. If M', the result of step 5, is less than the critical value of chi square, the hypothesis is *accepted* and further steps are disregarded.

Step 7. If the result of step 5 is greatly in excess of the critical value, the hypothesis is *rejected*. If it is but moderately in excess of the critical value, compute the correction, C. (In Table 12.23, subtract the reciprocal of the total of column 3 from the total of column 4. Multiply this result by the reciprocal of $3(k - 1)$, and add the result to 1.00.)

$$1.00 + (.1111)(.3747 - .0217) = 1.039$$

Step 8. Divide M', the result of step 5, by C, the result of step 7.

$$M = (5.30) \div (1.039.) = 5.10$$

Step 9. Refer to a table of chi square to test the significance of the M of step 8.

In the example of Table 12.23 the process would have been concluded with step 6 for $\alpha = .05$ or $\alpha = .01$ since, for d.f. $= 3$, the observed 5.30 is less than the 5 percent critical value, 7.82.

Usually educational variables are influenced by so many factors that it is necessary to employ more than two bases of classification in statistical analysis. Extensions of the methods of this chapter may be found in sources which deal with experimental design in detail (3, 7, 8, 13, 15). This chapter should be considered as only a brief introduction to analysis of variance and experimental design. Any serious planning of experiments should not be undertaken without a working knowledge of such topics as the "Latin Square" and "Analysis of Covariance," which have not been included in this book, nor without competent advice.

A large body of material is accumulating on *nonparametric* tests applicable to simpler designs such as those discussed in this chapter. These tests may be made with no assumptions regarding the distribution. Usually also they are simple to compute, but they are much less powerful than methods of this chapter in situations in which the assumptions are valid. Nonparametric methods often prove useful provided that (1) only a test of significance is required, and (2) the result of the nonparametric test is not close to the chosen critical value. References at the end of this chapter concerning such tests should be examined by the student planning extensive analysis (17, 18, 20, and 21). (See Ex. 18, Chapter 10, and Ex. 20, Chapter 12.)

EXERCISES

1. Assume that the four groups of students in the table of Appendix G are random samples from populations of students in the respective four high schools. Disregarding information concerning sex and college intentions, test the hypotheses that the 159 students grouped in the four high schools come from a single population with a single *mean* CTMM score. Test the hypothesis that these students come from a population with a single *variance*. What is the relationship between these two tests? Test the hypothesis that the four samples of physical science test scores for male students intending to go to college are from a single homogeneous population. Why would it not be advisable to attempt a two-way analysis of students classified by school and by college intentions?

2. By means of photographic apparatus several measures of eye-movement behavior of school children were made in a study of reading. One measure was the average duration per fixation in thirtieths of a second in reading prose. In the following table are the number of subjects, the means, and the standard deviations, s, for each of five grade levels included in the study:

Item	Grade				
	V	VI	VII	VIII	IX
Number of cases	58	65	39	48	34
Mean	7.99	7.83	7.91	7.82	7.34
Standard deviation	.94	.78	.61	.79	.77

Test the hypothesis, at the 1 percent level, of no difference among grade levels in average duration per fixation. Is it necessary to test homogeneity of variance in a problem with these results? What about normality? What difference would there be in results if the experiment had produced the same means and variances, but the sample sizes had been smaller, say, less than 10?

3. A state department of education was interested in the question of whether or not four different makes of school bus were equal in cost of operation. A random sample of three buses of each type was drawn from buses operating in each of three counties. Special cost accounting records were maintained on the 36 buses so that the mileage and operating cost, including depreciation, gas and oil, and maintenance and repair, could be recorded with reasonable accuracy for one year. The operating cost per mile for the 36 buses is reported in the table.

Make of Bus	County		
	Delaware	Schuyler	Putnam
A	.172, .343, .229	.286, .232, .191	.255, .192, .298
B	.213, .237, .451	.217, .486, .179	.414, .292, .169
C	.184, .521, .244	.602, .272, .166	.148, .396, .405
D	.198, .248, .302	.360, .260, .160	.329, .257, .218

By analysis of variance determine whether the data support the hypothesis of no difference in cost among the four types of bus. Describe the populations to which your statistical inference applies. What purpose was served in sampling from three different counties? In an experiment such as this should the experimenter strive for a large or small mean square *within*? What are the ways in which the experimenter may incorporate into his design the achievement of a larger or smaller *mean square within?* What are some of the factors not taken into account in the experiment which should account for some of the *mean squares within?* How could they be used to improve the design? In what manner would the design be improved? That is, would the conclusions be broader or narrower?

4. In a study of problem-solving behavior ten subjects each worked on problems A, B, and C. One measure of problem-solving behavior was the number of operations repeated by the subject. Results of this measure were:

Problem	Subject									
	1	2	3	4	5	6	7	8	9	10
A	1	2	1	2	1	5	1	5	6	4
B	6	0	6	2	4	4	3	9	4	8
C	1	5	5	1	5	4	8	7	4	6

At the 5 percent level test the hypothesis of no difference among the three problems and the hypothesis of no difference among subjects. What is your best guess as to the presence or absence of interaction between subject and problem? Is it possible to test for interaction in this type of problem? What would be the reliability of the three problems together as a single measure? What steps might be taken to increase the reliability coefficient if a measure of this type was sought?

5. What is the relationship of the sums of squares in the last row of Table 12.1 to the variances of the seven groups? What is the ratio of the highest variance to the lowest one? Does this suggest heterogeneity of variance of sufficient magnitude to vitiate the test of means? Check your answer to the previous question by means of the Bartlett test for homogeneity of variance.

6. Suppose that the scores for variables X_1 and X_2 of Appendix A were randomly assigned to the four double columns in which they appear. What would be the probability of finding a significant F ratio among means of either variable at the 5 percent level? Of finding significance among variances using the Bartlett test at the 5 percent level?

7. The following are two sets of single entries in rows and columns from two populations:

A				B		
5	7	10		9	17	10
9	11	14		14	11	9
12	14	17		5	8	12

In which population would you most expect to find interaction, A or B? Would it be possible to test for interaction in either A or B?

8. State in words a rule for finding the "pooled" variance of two or more groups.

9. When $n_1 = n_2$, prove that $(\bar{X}_1 - \bar{X})^2 + (\bar{X}_2 - \bar{X})^2 = \frac{1}{2}(\bar{X}_1 - \bar{X}_2)^2$.

10. Why is it sometimes desirable to use *interaction mean squares* as error in analysis of variance?

11. What components of variance are considered in analysis of variance with two-way classifications that are not present in analyses with a single basis of classification?

12. In a two-way classification design with replication, what population components are involved in the *residual mean squares?* The *interaction mean squares?* the *rows mean squares?* the *columns mean squares?*

13. Why is it not possible to test for interaction in a two-way table with a single entry in each cell?

14. What effect does an experimenter seek by introducing "control" factors? How is this effect accomplished?

15. In analysis of variance with more than one basis of classification why is it more fitting to use the term *mean square* than the term *variance* for ratios of sums of squares to degrees of freedom?

16. In testing differences between means of two groups, why is it that a two-tailed test is used with *t*, but a one-tailed test with *F?*

17. Why assume independence, normality, and homogeneity of variance in analysis of variance tests?

18. What are the reasons for including a basis of classification in an analysis of variance?

19. An experimental form of a test is administered to a group of subjects with time limits such that some complete all of the items of the test and others do not. A split-halves correlation is computed of scores on odd items versus scores on even items to determine reliability. What effect has the "timing" upon the correlation coefficient?

20. A random sample of eight classrooms was used in an experiment. In each classroom four methods of "stimulating class discussion," A, B, C, and D, were tried. For each classroom the four methods were ranked on a measure of "quality of discussion." The rankings were as follows:

Classroom	Method			
	A	B	C	D
a	1	2	4	3
b	3	2	4	1
c	2	1	3	4
d	1	3	2	4
e	1	2	3	4
f	3	1	2	4
g	2	1	4	3
h	2	1	3	4

A test described by Friedman (references 20 and 21) was used to test the hypothesis that the four methods are equally effective. The test statistic is

$$\chi_r^2 = \frac{12}{nk(k+1)} \Sigma T_r^2 - 3n(k+1)$$

where k is the number of treatments, n is the number of replications, and T_r is a rank total for a replication. This statistic is approximately distributed as χ^2 with $(k - 1)$ degrees of freedom. At the .05 level test the hypothesis that there is no difference in the four treatments.

REFERENCES

1. Binder, Arnold, "The Choice of an Error Term in Analysis of Variance Designs," *Psychometrika*, 20: 29–50, March 1955.
2. Cochran, William G., "Some Consequences When the Assumptions for the Analysis of Variance are Not Satisfied," *Biometrics*, 3: 22–38, March 1947.
3. Cochran, William G., and Gertrude M. Cox, *Experimental Designs*, New York, John Wiley and Sons, 1950, Chapters 1, 2, and 3.
4. Dixon, Wilfrid J., and Frank J. Massey, *Introduction to Statistical Analysis*, New York, McGraw-Hill Book Co., 1951, Chapters 10, 14, and 17.
5. Eisenhart, Churchill, "The Assumptions Underlying the Analysis of Variance," *Biometrics*, 3: 1–21, March 1947.
6. Eisenhart, Churchill, Millard W. Hastay, and W. Allen Wallis, *Techniques of Statistical Analysis*, New York, McGraw-Hill Book Co., 1947, Chapter 15.
7. Fisher, Ronald A., *The Design of Experiments*, Sixth Ed., New York, Hafner Publishing Co., 1951.
8. Fisher, Ronald A., *Statistical Methods for Research Workers*, Eleventh Ed., New York, Hafner Publishing Co., 1950, Chapters 7 and 8.
9. Gourlay, Neil, "F-test Bias for Experimental Designs in Educational Research," *Psychometrika*, 20 : 227–48, September 1955.
10. Hartley, Herman O., "The Maximum F-Ratio as a Short-Cut Test for Heterogeneity of Variance," *Biometrika*, 37: 308–12, December 1950.
11. Hoyt, Cyril "Test Reliability Estimated by Analysis of Variance," *Psychometrika*, 6: 153–60, June 1941.
12. Jackson, Robert W. B., *Application of the Analysis of Variance and Covariance Method to Educational Problems*, Toronto, University of Toronto, Department of Educational Research Bulletin, No. 11, 1940.
13. Kempthorne, Oscar, *The Design and Analysis of Experiments*, New York, John Wiley and Sons, 1952.
14. Kogan, Leonard S., "Variance Designs in Psychological Research," *Psychological Bulletin*, 50: 1–40, January 1953.
15. Lindquist, Everet F., *Design and Analysis of Experiments in Psychology and Education*, Boston, Houghton Mifflin Co., 1953.
16. Lindquist, Everet F., editor. *Educational Measurement*, Washington, D.C., American Council on Education, 1951. Robert L. Thorndike, "Reliability," pp. 560–620.
17. Mood, Alexander M., *Introduction to the Theory of Statistics*, New York, McGraw-Hill Book Co., 1950, Chapter 16.
18. Savage, I. Richard, "Bibliography of Nonparametric Statistics and Related Topics," *Journal of the American Statistical Association*, 48: 844–906, December 1953.
19. Snedecor, George W., *Statistical Methods*, Fourth Ed., Ames, Iowa, The Collegiate Press, 1946, Chapters 10 and 11.
20. Walker, Helen M., and Joseph Lev, *Statistical Inference*, New York, Henry Holt and Co., 1953, Chapters 8 and 9.
21. Wilcoxon, Frank, *Some Rapid Approximate Statistical Procedures*, New York, American Cyanamid Co., 1949.

CHAPTER 13

Special Correlation Methods

The investigator interested in determining the correlation between two variables must often work with data not in the form discussed in Chapter 8 for which the Pearson-product-moment-coefficient is appropriate. There are several situations which occur of this type for which there are special correlation coefficients similar to or approximating the Pearson coefficient and which are in frequent use in educational research.

13.1 THE CORRELATION INDEX AND CURVILINEAR CORRELATION

We saw in Section 8.6 that the square of the correlation coefficient may be interpreted as the ratio of the variance of regression values to total variance in the dependent variable. Also, in terms of *residual* variance, we may write from equation 8.27 the relationship

$$\rho_{12}^2 = 1 - \frac{\sigma_{1.2}^2}{\sigma_1^2} \tag{13.1}$$

It will be recalled that in simple correlation the residual variance is measured from one of the regression lines, a *straight line*. By a slight modification of equation 13.1 we may derive a formula called the *correlation index*, based on the residual variance of a measure X_1 from any predictor of it, X_1'. The predictor X_1' may be a function of one or several other measures, and it need not be linear. Computation of the correlation index is simplest in terms of *sums of squares*, thus

$$R^2 = 1 - \Sigma(X_1 - X_1')^2/\Sigma(X_1 - \bar{X}_1)^2 \tag{13.2}$$

where

$$\Sigma X_1 = \Sigma X_1'$$

It is possible by the method of least squares to fit a variety of nonlinear regression curves to bivariate distributions—parabolas, circles, exponential functions, etc.—which might predict X_1 from X_2 or vice versa.

Such methods are beyond the scope of this book. When this is done, however, equation 13.2 will yield a measure of the degree of correlation.

13.2 THE CORRELATION RATIOS

A fairly simple measure of nonlinear correlation may be computed for bivariate data which are grouped as in the correlation table of Section 8.7. The fitting of curves in order to obtain predicted values X_1' is not required. This measure, the *correlation ratio*, is obtained by measuring residuals or "errors of estimate" from the means of columns (or rows, as the case may be).

There are two ratios, one in terms of residuals of X_1 computed from means of *columns*, and the other in terms of residuals of X_2 computed from means of *rows*. These may be defined as

$$\eta_{12}^2 = 1 - \frac{\sigma_1^2(w)}{\sigma_1^2}$$

and

$$\eta_{21}^2 = 1 - \frac{\sigma_2^2(w)}{\sigma_2^2} \tag{13.3}$$

where $\sigma_1^2(w)$ is the variance *within* columns, and $\sigma_2^2(w)$ is the variance *within* rows.

The actual computation of *eta* for either variable is best accomplished by means of sums of squares. The square of *eta* is 1 minus the ratio of the within sum of squares to the total sum of squares of the predicted variable. Thus, in the notation of Chapter 12,

$$\eta^2 = 1 - \frac{\sum_i\sum_j(X_{ij} - \bar{X}_i)^2}{\sum_i\sum_j(X_{ij} - \bar{X})^2} \tag{13.4}$$

where X_{ij} is the jth individual in the ith column (or row), $\bar{X}_i$ is the mean of the ith column (or row), $\bar{X}$ is the general mean, n_i is the frequency in the ith column (or row), and summations are over all k columns (or rows).

We may combine terms of the right-hand side of equation 13.4, and, using our knowledge that the *total* sum of squares minus the *within* sum of squares is equal to the *between* sum of squares, derive an alternative formula for computing the square of *eta*:

$$\eta^2 = \frac{\sum_i n_i(\bar{X}_i - \bar{X})^2}{\sum_i\sum_j(X_{ij} - \bar{X})^2} \tag{13.5}$$

In this form, the square of *eta* is the ratio of the sum of squares of deviations (weighted by the number of cases in each column) of column means from the general mean to the total sum of squares.

The procedures are illustrated with the data of Table 13.1. The mean of each column is computed. These means, plotted on the chart of Fig. 13.1 and joined by lines, are seen to form a broken line. The purpose of computing *eta* is, in effect, that of finding out how closely individual pairs of measures cluster about this line.

TABLE 13.1

ANNUAL COST OF TRANSPORTATION PER PUPIL TRANSPORTED AND NUMBER OF PUPILS PER SQUARE MILE IN 87 SCHOOL DISTRICTS

Average cost per Pupil (X_1)	Pupils per Square Mile (X_2)							Total
	1.0–1.6	1.7–2.3	2.4–3.0	3.1–3.7	3.8–4.4	4.5–5.1	5.2–5.8	
90–100	1	3	2					6
79–89		3	4	2				9
68–78	2	7	7	3	1	1		21
57–67	2	4	8	4	3	3	1	25
46–56		3	11	2	1	2	1	20
35–45			2	1	2			5
24–34					1			1
Total	5	20	34	12	8	6	2	87

TABLE 13.2

COMPUTATION OF ETA

X_2	n_i	$\sum^{n_i} X$	$\bar{X}_i$	$(\bar{X}_i - \bar{X})$	$n_i(\bar{X}_i - \bar{X})^2$
(1)	(2)	(3)	(4)	(5)	(6)
5.5	2	113	56.50	− 8.53	145.52
4.8	6	361	60.17	− 4.86	141.72
4.1	8	419	52.38	−12.65	1,280.18
3.4	12	777	64.75	− .28	.94
2.7	34	2,174	63.94	− 1.09	40.40
2.0	20	1,449	72.45	+ 7.42	1,101.13
1.3	5	365	73.00	+ 7.97	317.60
All columns	87	5,658	65.03	—	3,027.49

By various methods already familiar to us we may find the total sum of squares, $\sum^{k}\sum^{n_i}(X_{ij} - \bar{X})^2$, for the variable cost per pupil to be 18,559. The means of columns and the deviations of means of columns are shown in Table 13.2. The squares of the latter, weighted by n_i, are shown in column 6, the total of which yields the *between means* sum of squares, 3,027. This may be substituted in equation 13.5 to find $\eta_{21}^2 = 3,027/18,559 = .1631$, and $\eta_{12} = .404$.

Although we may note from Fig. 13.1 that the correlation is negative, our method of computing *eta* disregards sign and is always positive.

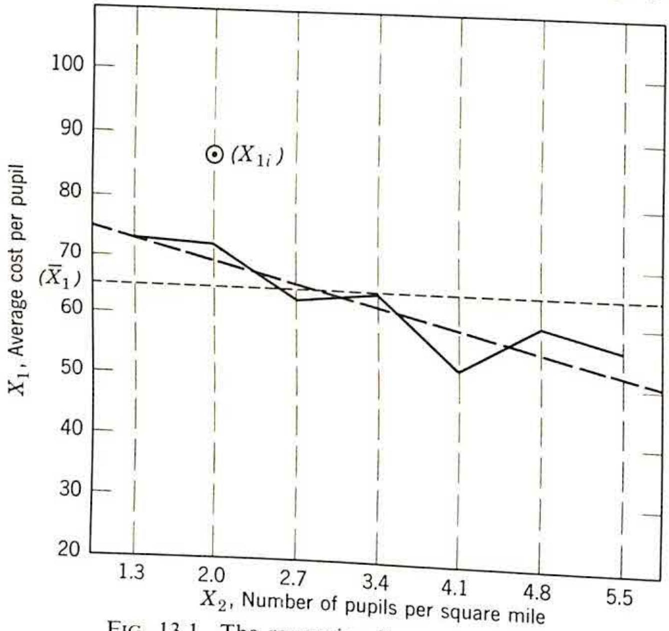


FIG. 13.1. The regression line and line of means for columns of Table 13.1

A more accurate partitioning of sums of squares will show the *between* to be 3,029 and the *within* 15,530. The latter may be used in equation 13.4 to compute *eta*. A similar procedure would be followed in finding η_{21}, the other *correlation ratio*, by computing the sums of squares of means of *rows* and *total* sum of squares for pupils per square mile.

13.3 THE SIGNIFICANCE OF ETA

The methods of analysis of variance of Chapter 12 may be used directly to test the hypothesis $H : \eta = 0$. In Table 13.3 is such an analysis. An

F ratio, using the within columns (or residual) mean square as "error," shows the variance between means to be significant at the 5 percent level. The correlation ratio, η_{12}, a function of the *between means* variance, is thus also significant at the 5 percent level.

TABLE 13.3

Analysis of Variance for Eta

Source of Variation	Degrees of Freedom	Sum of Squares	Mean square
Total	86	18,559	—
Between column means	6	3,029	505*
Within columns	80	15,530	194

* $F = 505/194 = 2.60$.

From the relationship of η and sums of squares given in equations 13.4 and 13.5, it is easily shown that the F ratio used in making this test is given by

$$F = \frac{\eta^2/(k-1)}{(1-\eta^2)/(N-k)} \tag{13.6}$$

13.4 TESTING LINEARITY OF REGRESSION

An extension of the variance analysis of transportation costs in Table 13.1 will permit a test of the departure of the line of means from linearity. The basis of comparison is the appropriate regression line. In our example, an application of the methods of Section 8.7 will show that $r = -.339$, and that the regression equation is $\hat{X}_1 = -5.16X_2 + 79.95$. The latter is drawn in dashes in Fig. 13.1. Also shown as a dotted line in Fig. 13.1 is the mean, $\bar{X}_1$, from which deviations are measured to determine the *total* sum of squares.

We observe that r_{12} is less than η_{12}. This is because the residual variance (within columns) from our computation of *eta* is taken from column *means*. As seen in Section 4.3, sums of squares of deviations are *least* when measured from a mean. In other words, the data conform more closely (that is, have *less* residual variance) to the column means than to straight line. Therefore, $\eta_{12}^2 \geq r_{12}^2$, and $\eta_{21}^2 \geq r_{12}^2$. The correlation ratio and the correlation coefficient will be the same only if all the means lie exactly on the regression line. Of course, this never happens.

Because of the vagaries of sampling, column means will not lie along a straight line even for a population for which linearity is the true relation.

The sum of squares of deviations of column means from the regression line is equal to the difference between (1) the sum of squares *within columns*, and (2) the sum of squares of *residuals* from the regression line. A variance computed from this sum of squares of deviations of column means from regression may be used to test the difference between η and r, that is, the hypothesis that the sample comes from a universe whose regression (that is, correlation) is linear. The corresponding comparison of η with r for rows may sometimes be made, and it should be noted that one regression can be linear and the other nonlinear, though not in a bivariate normal distribution.

An examination of Fig. 13.1 will show that the deviation of any individual score, X_{1i}, from the general mean, $\bar{X}$, consists of three parts:

(*a*) The deviation of the regression value from $\bar{X}$,

(*b*) The deviation of the column mean from the regression value,

(*c*) The deviation of the individual score from the column mean.

The sum of squares of all these deviations equals the "total" sum of squares in X_1. We show these sums of squares in Table 13.4 and proceed to explain how they are derived.

We already have the last of the three components of sums of squares for our example. It is the *within columns* sum shown in Table 13.3. From computations for the correlation table for Table 13.1, the following quantities were found: $\Sigma x_1x_2 = -412.9$ and $\Sigma x_2^2 = 80.06$. Substituting in equation 8.22, we find the regression sum of squares to be 2,129. This is entered in Table 13.4 for the variance component (*a*) with 1 d.f. The

TABLE 13.4

Analysis of Variance for Testing Linearity of Regression

Source of Variation	For Testing	Degrees of Freedom	Sum of Squares	Mean Square
Total	—	86	18,559	—
(*a*) Deviation of predicted values from general mean	Significance of regression and correlation	1	2,129	2,129**
(*b*) Deviation of column means from regression	Departure from linearity of regression	5	900	180
(*c*) Within columns	(*a*)	80	15,530	194

(*a*) Used as error mean square, that is, denominator in *F* ratio.

** In general use, double asterisk indicates significance at the 1 percent level.

difference between this sum and the *between means* sum of Table 13.3 is used for the (*b*) component, the sum of squares of departures of means from regression, with 5 d.f.

As previously, the (*c*) variance is the error variance. We test the variance (*b*) by means of the F ratio, 180/194. This is a little less than 1.00. Therefore, the departures of means from regression are not great enough to justify rejecting the hypothesis of linearity of regression of X_1 on X_2. A similar test of linearity of regression of X_2 on X_1 may be made if desired.

The general scheme for analyses such as Table 13.4, with k arrays (columns or rows), is:

Component	Degrees of Freedom
Total	$N-1$
(*a*) Regression	1
(*b*) Departure from linearity	$k-2$
(*c*) Within arrays	$N-k$

Note that the sum of squares (*b*) plus the sum of squares (*c*) equal $\Sigma x_{1.2}^2$, with $(N-2)$ degrees of freedom. In this section our objective was to test *linearity* of regression, but the analysis may also be used for testing the *significance* of regression or correlation. From Table 13.4 we see that the residual variance made up of these two components would be $16{,}430/85 = 193.3$. This is an appropriate error variance for testing the regression variance, component (*a*), provided that regression is found to be linear. We compute $F = 2{,}129/193.3 = 11.01$ for 1 and 85 d.f. This is significant at the 1 percent level. Therefore, the hypothesis $H: \rho = 0$, or $H: B_{12} = 0$, is rejected. It can be shown that, when regression is linear, F is given by

$$F = \frac{r^2}{(1-r^2)/(N-2)} \tag{13.7}$$

This is precisely the square of the t test of Section 9.4 and will give results identical to the significance tests of that section.

13.5 RANK CORRELATION

Frequently it is not possible to obtain pairs of measures on individuals, but they can be ranked. In Table 13.5 are shown the rankings by two teachers of twelve pupils on a characteristic of "initiative." A fairly simple device for computing correlation of such ranks is the *Spearman rank-correlation coefficient*. It was used originally as an estimate of the Pearson-product-moment-coefficient and more recently as a test of independence. It can be used as a simple method of computing correlation for small samples, or for testing the independence of ranks. It has the advantage that no assumption is made concerning the distribution of the two measures. It is one of the nonparametric tests mentioned at the end of Chapter 12.

TABLE 13.5

Correlation of Rankings of Twelve Pupils by Two Teachers

Individual	Teacher		d	d^2
	I	II		
A	1	2	-1	1
B	2	5	-3	9
C	3	6	-3	9
D	4	1	$+3$	9
E	5	4	$+1$	1
F	6	9	-3	9
G	7	7	0	0
H	8	3	$+5$	25
I	9	11	-2	4
J	10	12	-2	4
K	11	8	$+3$	9
L	12	10	$+2$	4
Total	—	—	—	84

If data on pairs of measures are not in rank form, they can be ordered on available measures and ranks assigned. The usual procedure in case of ties in ranks is to assign to each of the individuals tied the average of the rank values they would receive had they not been tied. For instance, if three individuals tie for seventh place, they would receive each the rank of eighth, the average of the ranks 7, 8, and 9. Four individuals tying for seventh place would each receive a rank of 8.5. Converting *measures* to *ranks* may be necessary before the *rank-difference correlation* method is to be used. It is applicable to any population, not just the normal population.

The formula for the *rank-correlation* coefficient is as follows:

$$r' = 1 - \frac{6\Sigma d^2}{n(n^2 - 1)} \tag{13.8}$$

where d is the difference between an individual pair of ranks.

In the last two columns of Table 13.5 appear the rank differences and the squares of the rank differences. The sum of these squares is found to be 84. The rank-correlation coefficient is, therefore, computed to be $1 - (504)/(1716) = .706$.

Tables for testing the hypothesis $H : r' = 0$, or its equivalent, $H : \Sigma d^2 = n(n^2 - 1)/6$ are available (2, 4, and 5). Reference to these tables shows that the observed rank-correlation coefficient of .706 for Table 13.5 is significant at the 5 percent level.

13.6 BISERIAL CORRELATION

Frequently one of two measures to be correlated is reported in a dichotomy. For example, in one correlation problem, one variable was *amount of education* of a group of adults. The survey supplied only information on who had graduated from college and who had not. There was thus a measure of extent of education of individuals only in terms of whether or not they had graduated from college. Of course, amount of schooling is a continuous variable. We can imagine a very elaborate measure of years or months or hours of formal education which would produce a *continuous* distribution.

When one variable is a dichotomy and the other a continuous variable, there are two different coefficients of correlation which can be used. The first of these to be considered is the *biserial* correlation coefficient. It is an estimate of the correlation which would be obtained if *both* measures were available in *continuous* graduated form. It is derived directly from the basic formula for the Pearson-product-moment-coefficient. Therefore, its use involves the assumptions of the correlation coefficient, and it is of doubtful value if the continuous variable is not approximately normally distributed and if it is unreasonable to assume a normal distribution of the dichotomized trait. Although linearity of regression is assumed, no regression formula may be based upon the biserial r and there is not a satisfactory error of estimate.

We will illustrate the application of biserial r. Suppose that a group of 114 students were given a paper-and-pencil test of knowledge about undertaking some task. This might be, for instance, repairing a radio set. Suppose, furthermore, that it is desired to compare such scores with the actual performance of the same subjects on a performance test which involves finding the defective component in a real radio with actual test equipment. One result of a single performance test item could be a simple "pass-fail," classing those who find the defective component as "pass," and those who do not as "fail." The dichotomy for a test of several such items may be determined by a specified passing score. The results could be tabulated as in Table 13.6.

TABLE 13.6

FREQUENCY DISTRIBUTION OF PAPER-PENCIL TEST SCORES, GROUPS FAILING AND PASSING PERFORMANCE TEST

X	Success on Performance Test		Total
	Pass	Fail	
33–35	2		2
30–32	4	1	5
27–29	8	2	10
24–26	9	2	11
21–23	12	5	17
18–20	15	8	23
15–17	9	12	21
12–14	4	8	12
9–11	3	5	8
6– 8	1	2	3
3– 5		1	1
0– 2		1	1
Total	67	47	114

The *biserial* correlation coefficient is given by

$$r_{bis} = \frac{\bar{X}_p - \bar{X}_q}{s_x} \cdot \frac{pq}{f(X)} \tag{13.9}$$

where $\bar{X}_p$ is the mean of the variable X for one of the two groups of the dichotomized variable, $\bar{X}_q$ is the mean of the other group, p is the proportion of cases in the first group, $q = (1 - p)$ is the proportion of cases in the second group, s_x is the standard deviation of the X variable in both groups, and $f(X)$ is the ordinate of the unit normal curve at the point which cuts off p proportion of cases on one tail and q proportion of cases on the other. Note that if $\bar{X}_q$ is larger than $\bar{X}_p$, the coefficient is negative. It is wise to designate the p and the q groups so that the correlation will have a sign consistent with the logical order of the dichotomies. For some dichotomies there is, of course, no logical basis for ordering the two classes.

By the usual methods we compute $\bar{X}_p = 21.2$; $\bar{X}_q = 16.3$; and $s = 6.47$. We note also that $p = 67/114 = .59$ and $q = 47/114 = .41$. We can find the ordinate $f(X)$ for which $\int_0^x = .09$ from the table in Appendix C. This is found to be .389. In using the table in Appendix C it is to be recalled that the areas given are measured from the *mean* and not from either tail of the distribution. For this reason the table is entered for the area .09, not .59. Substituting these values in formula 13.9, we find $r_{bis} = (4.9/6.47)(.24/.389) = .48$.

From the relationship of the mean of both dichotomized groups to the mean of the two groups separately, $\bar{X} = p\bar{X}_p + q\bar{X}_q$, and the relationship $p + q = 1$, it is possible to express the formula for biserial r in terms of the mean of one of the groups and the general mean as follows:

$$r_{bis} = \frac{\bar{X}_p - \bar{X}}{s_x} \cdot \frac{p}{f(X)} \tag{13.10}$$

This equation is more economical when a series of dichotomized measures are being correlated with a single continuous measure. Unusual results may be obtained in computing the biserial correlation coefficient. It is possible to have distributions for which the biserial correlation is more than unity.

Should the trait represented by the dichotomous variable be one which cannot be considered continuous, it is necessary to use the *point biserial* correlation. It is often a matter of psychological definition whether a dichotomy represents a *continuous* function or whether it represents merely *two points*. For instance, one dichotomous measure might be based on whether students continue or drop out of school or college. Literally a student is either retained on the rolls or he is no longer in attendance. A clean-cut distinction, therefore, exists between the two categories in which an individual may be placed. On the other hand, staying in school or not staying in school may be considered a continuum since there are many who stay in attendance who possess traits of low scholarship, unacceptable behavior, low economic status, or the like in degrees close to the point where they would have to leave school. Also, many who left school did so much more conclusively than did others. Potential for staying in or not staying in school may thus be viewed as a continuous variable. Similarly, passing or failing a test item may be an objectively determined, clean-cut dichotomy. Yet the underlying skill or ability measured may be continuous, and it can be argued that among those who pass there is a great variation in the ease with which they pass, just as there is a great variation in the degree of failure existing among those who fail the item.

If the dichotomous variable represents two points, then the *point biserial* coefficient should be computed, using one of the following two formulas:

$$r_{p.\,bis} = \frac{\bar{X}_p - \bar{X}_q}{\sigma_x} \sqrt{pq} \qquad (13.11)$$

$$r_{p.\,bis} = \frac{\bar{X}_p - \bar{X}}{\sigma_x} \sqrt{p/q} \qquad (13.12)$$

Letting the two performance categories of Table 13.6 represent two points, we use equation 13.11 to find that $r_{p.\,bis} = .37$.

If the two classes of the dichotomous variable are assigned numbers such as 1 and 0, the usual formulas for the product-moment correlation coefficient will give the same result as the *point biserial* coefficient. The point biserial coefficient is thus a special case of the Pearson coefficient.

The sampling distribution of the *biserial* coefficient is not known, but in some cases approximate confidence intervals can be obtained by a modification of Fisher's z transformation (reference 7).

It should be noted that the data of Table 13.6 may be viewed as a situation requiring the comparison of two sample means. Where the appropriate assumptions hold, equation 11.15 may be used as a t test for the difference between the two means. This is equivalent to testing the significance of the *point biserial* coefficient, but *not* the *biserial* coefficient. The t of this test may be written

$$t = r_{p.\,bis} \frac{\sqrt{N-2}}{\sqrt{1 - r^2_{p.\,bis}}} \qquad (13.13)$$

This is the same as the t of equation 9.2 and is used to test the hypothesis, $H : \rho_{p.\,bis} = 0$.

Confidence intervals can be computed for the *point biserial* coefficient by means of special tables (6, 7).

13.7 THE CONTINGENCY COEFFICIENT

When two variables are classified in categories, especially in categories which are not ordered, the *coefficient of mean-square contingency* is sometimes used. It is based upon χ^2. In Chapter 10 we used χ^2 to test *independence*, that is, to find out whether or not there was association between variables arranged in some system of classification.

The *contingency coefficient* is

$$C = \sqrt{\frac{\chi^2}{n + \chi^2}} \qquad (13.14)$$

In Section 10.6 we found a χ^2 measuring the "absence of independence" of economic class in responses to a question in a survey to be 5.69. Substituting in equation 13.14, where $n_{.} = 125$, we find $C = .209$.

The possible range of the contingency coefficient is from zero to one. Therefore, the relationship does not appear to be a high one. In fact, it is not significantly different from zero, as shown by the chi-square test of Section 10.6.

The coefficient is always considered to be positive, but this presents no difficulty in problems of finding the association between nonorderable variables such as color of hair versus color of eye, college attendance versus undergraduate major, and the like.

An advantage of this measure of relationship is that there need be no assumption concerning the nature of underlying distributions. It is particularly suited to data which are classifiable truly in "point" distributions, but it is not directly interpretable as the Pearson-product-moment-coefficient. The maximum possible value of C depends upon the number of categories used in the contingency table. Therefore, coefficients from tables with varying numbers of categories are not comparable. In a 2 × 2 table, the maximum value of C is .707, whereas in a 10 × 10 table the maximum value is .949.

Karl Pearson has shown that the contingency coefficient would be the same as the correlation coefficient if variables were continuous, if the distributions were normal, if the regression was linear, and if the number of categories was sufficiently large. Corrections have been developed for differences in number of categories in a table from which the contingency coefficient is computed so that results are comparable to r. This is rarely useful, both because of the complexity of adjustments necessary and the difficulty of satisfying the assumptions in problems for which the contingency coefficient is most suitable.

13.8 FOURFOLD CORRELATION—PHI

There is one other measure of correlation which is related to chi square and is applicable in 2 × 2 tables when both variables are true dichotomies. This is the *phi coefficient*, ϕ. An advantage over the coefficient of contingency is that the phi coefficient can be shown to be the Pearson correlation coefficient for two-point distributions. A formula for phi in terms of chi square is:

$$\phi = \sqrt{\chi^2/n} \qquad (13.15)$$

where chi square is computed from a 2 × 2 table *without* correction for continuity.

In the exercise of Section 10.10 we computed chi square for the fourfold table involving 35 students classified as to college attendance intentions and as to sex. Eliminating the correction term, $\frac{1}{2}n$, we recompute the chi square of that section and find it to be 2.31. Applying formula 13.15, we find that $\phi = \sqrt{(2.31)/(35)} = .257$.

The same result may be achieved by computing phi directly from the formula

$$\phi = \frac{ad - bc}{\sqrt{(a + b)(a + c)(b + d)(c + d)}} \tag{13.16}$$

In the foregoing example, using this equation, $\phi = 77/299.94 = .257$.

The value of the phi coefficient tends to be less than that of the tetrachoric coefficient (explained below). Phi is influenced by the choice of cut-off point dichotomizing the two variables. It is best to avoid it when either dichotomy is split into extremely unequal categories, say, 10 to 90 percent. The limits of phi are -1.0 and $+1.0$. The assignment of letters to cells in the fourfold table should be made so that the numerator of (13.16) will yield the sign appropriate to the interpretation. Phi is useful for computing the intercorrelation of dichotomously scored test items.

13.9 TETRACHORIC CORRELATION

Another coefficient, the *tetrachoric correlation coefficient*, may be used with two "continuous variables," both of which have been dichomotized. It is derived from bivariate normal correlation theory. Therefore, the use of this coefficient implies the assumption that there exist normally distributed continuous underlying variables even though these have not been observed. For instance, the heights and weights of individuals in a population may conceivably meet these assumptions, but the information available to an investigator for these two measures may be only a double dichotomy. He may have sample data for the number over and under 6 feet, for those weighing less, and for those weighing more than 175 pounds. His information would thus be only a double dichotomy in these variables, though they might actually be in a bivariate normal distribution, thus satisfying also the conditions of linearity and homoscedasticity.

The tetrachoric coefficient is an estimate of the product-moment correlation. It is used in test construction, where it is desired to *correlate* test items (assumed to represent a continuous underlying trait) which are measured on a pass-fail basis with some criterion which also is dichotomized as "success-failure."

Direct computation of the tetrachoric coefficient is almost prohibitively complicated, since it involves many powers of the coefficient. The equation used, to four terms only, is

$$\frac{bc - ad}{n^2 y_1 y_2} = r_t + \frac{z_1 z_2 r_t^2}{2!} + \frac{(z_1^2 - 1)(z_2^2 - 1)r_t^3}{3!} + \frac{(z_1^3 - 3z_1)(z_2^3 - 3z_2)r_t^4}{4!} + \cdots \tag{13.17}$$

Sometimes only the first two terms on the right-hand side are used, thus avoiding powers of r_t higher than the second. This permits a relatively simple approximate solution involving only a quadratic equation. The terms involving higher powers of r are usually quite small.

The chief purpose of presenting equation 13.17 is to permit the reader to see how directly it is concerned with the assumption of normality. In this equation, a, b, c, and d are the frequencies in the four cells of the fourfold table, and n is the total number of cases in the distribution. The other variables come directly from the normal distribution. A table of areas, ordinates, and deviates, such as that in Appendix C, is used. The proportion of cases in each category for each variable is considered a proportion of the area of the normal curve. A table such as that of Appendix C is entered to find the ordinates y_1 and y_2 at the points of division of the normal curve into the two dichotomies. Similarly, the normal deviates, z_1 and z_2, are found from the normal table. For instance, if the variable X_1 was dichotomized so that 55 percent of the cases were in the high category and 45 percent in the other, the point of division in the normal curve would be at $z_1 = +.126$, and the corresponding value of y_1 would be .396.

Because of interest in the use of tetrachoric correlation when a large number of coefficients are required, such as in test item analysis, computing aids have been developed. The Thurstone computing diagrams are very suitable for this purpose and yield two-place accuracy (reference 1).

The sampling error of the tetrachoric coefficient is considerably greater than that of the Pearson coefficient. Much information is lost if data are reduced to a fourfold table. For this reason, as with other measures of fourfold correlation, it should be avoided particularly when samples are small. Where samples are quite large, as in some programs of test construction, the loss in reliability is partly offset by the gain in computation time possible, especially when computing diagrams are used.

As with other estimates of correlation from fourfold tables, care must be exercised in interpreting results obtained from unusually one-sided dichotomies. The nearer the dichotomies split the sample 50-50, the more reliable they are.

13.10 COMPARING MORE THAN TWO INDEPENDENT CORRELATIONS

Fisher's z transformation and chi square may be used to test the hypothesis that several samples come from a single population with a correlation ρ. The hypothesis may be stated

$$H : \rho_1 = \rho_2 = \rho_3 = \cdots = \rho_k$$

when there are k observed values of r.

The justification for the technique is most easily understood for the case of k samples of equal sample size. First we transform the r's to z''s. If the hypothesis is true, the observed values of z' are values from samples of size n from populations with a common z'. In Section 9.5 it was noted that the distribution of sample values of z' is approximately normal.

Next we recall from equation 10.9 that the sum of squares of the unit normal deviate is distributed as chi square. Hence, if our k samples are of equal size, we could:

(1) Compute $\bar{z}'$, the mean of the z''s.

(2) Find the sum of squares of deviations of observed z''s from $\bar{z}'$.

(3) Compute σ_z^2, from equation 9.5

(4) Find the ratio of step 2 to step 3.

The result of step 4, as indicated by equation 10.9 would be a statistic distributed as chi square. If this ratio is sufficiently large for $(k - 1)$ degrees of freedom, the probability of its occurrence will be small and will cast doubt on the hypothesis.

Since the k samples may not be of equal size, the procedure is modified slightly by appropriate "weights" to take into account differences in sample size. The following describes the more general procedure, the rationale of which is similar to the above discussion.

As before, the first step is to transform the r's to z''s. Weighting each one by the reciprocal of its variance, $(n_i - 3)$,

$$\bar{z}' = \frac{\Sigma(n_i - 3)z_i'}{\Sigma(n_i - 3)} \tag{13.18}$$

The *weighted* sum of squares of deviations of the z''s from $\bar{z}'$ is

$$\Sigma(n_i - 3)(z' - \bar{z}')^2 \tag{13.19}$$

Weighting the squared deviations by the reciprocal of the corresponding variances is the same as dividing each by the corresponding variance. The result is equivalent to expressing each deviation in standard form.

Thus from equation 10.9 we see that equation 13.19 is distributed as χ^2 with d.f. $= (k - 1)$.

Computation is simplified by using the formula

$$\chi^2 = \Sigma(n_i - 3)(z')^2 - [\Sigma(n_i - 3)z']^2/\Sigma(n_i - 3); \quad \text{d.f.} = k - 1 \quad (13.20)$$

which is simply equation 4.14 generalized to include weights.

Obviously, if all the ρ's are equal, their transformations would be also equal. Therefore, we test the hypothesis of no difference among correlations by examining the extent of variation of the z''s, using equation 13.20.

Table 13.7 gives the correlations between (*a*) *amount* of teacher participation in policy making and (*b*) degree of *responsibility* for participating in policy making, from samples of teachers from four school systems. Substituting in equation 13.20, we have

TABLE 13.7

COMPARING SAMPLE CORRELATIONS FROM FOUR SCHOOL SYSTEMS

School System	n_i	r_i	$(n_i - 3)$	z'_i	$(n_i - 3)z'_i$	$(n_i - 3)(z'_i)^2$
A	14	.73	11	.929	10.219	9.493
B	16	.48	13	.523	6.799	3.556
C	14	.13	11	.131	1.441	.189
D	11	.56	8	.633	5.064	3.206
Total	55	—	43	—	23,523	16.444

$$\chi^2 = 16.444 - 12.868 = 3.58$$

This is not large enough to be significant at the 5 percent level with 3 d.f.

Since the test does not give us grounds for rejecting the hypothesis, we may be justified in considering the four samples as if from a common population correlation and average the correlations by first averaging the corresponding z transformations, using equation 13.18 and converting the average z' to the corresponding r. Substituting in equation 13.18 from Table 13.7, we have

$$\bar{z}' = 23.523/43 = .55$$

The corresponding value of r is found in Appendix F to be .50. The weighted mean, $\bar{z}'$, is approximately normally distributed with variance

$$\sigma^2_{\bar{z}'} = 1/\Sigma(n_i - 3) \quad (13.21)$$

This permits tests of hypotheses concerning the population value of $\bar{z}'$ (and the corresponding average ρ). In the example, $\sigma^2_{\bar{z}'} = 1/43 = .0233$. The square root is $\sigma_{\bar{z}'} = .15$. Approximate 95 percent confidence limits for $\bar{z}'$ would be $[.55 - (1.96)(.15)]$ and $[.55 + (1.96)(.15)]$ or .26 and .84. Converting confidence limits of $\bar{z}'$ to corresponding limits for ρ, $P(.25 < \rho < .69) = .95$.

This example is of further interest in revealing the limited usefulness of correlations based upon small samples. There is a wide variation in the four correlations of Table 13.7, yet we find it quite reasonable to expect them to occur by random sampling alone from a common population correlation. Even the average correlation for the four groups based upon all fifty-five individuals has a wide 95 percent confidence band of .44.

The approximate methods of this section are affected by a positive error in computing each z' which is roughly in the magnitude of $\rho/2(n_i - 1)$. This is of some importance when n_i is as small as in the samples in Table 13.7 and k is large, since the error is greater, the smaller the sample size, and in equation 13.18 the errors accumulate.

EXERCISES

1. Describe the types of correlation problems for which each of the following measures of correlation is appropriate: (*a*) The correlation ratio. (*b*) Rank correlation. (*c*) Biserial correlation. (*d*) Point biserial correlation. (*e*) Coefficient of mean-square contingency. (*f*) Phi coefficient. (*g*) Tetrachoric correlation. (*h*) Correlation index.

2. The data in the following table are figures on annual current expense per pupil in a sample of fifty-three high schools grouped into five categories according to numbers of teachers.

Number of Teachers	Expenditure per Pupil
0–10	738, 345, 252, 590, 473, 433, 420, 290, 435, 588, 264, 705, 419, 355, 345
11–20	225, 337, 247, 313, 591, 425, 304, 285, 355, 400, 222, 400
21–30	378, 163, 231, 217, 224, 245, 462, 280, 260
31–40	259, 362, 310, 287, 262, 374, 362, 205, 355
41–50	376, 293, 259, 307, 377, 209, 221, 255

(*a*) Find the regression coefficient for predicting expenditure from size as measured by number of teachers.

(*b*) Compute the appropriate correlation ratio for examining variations in expenditure among high schools of different size.

(*c*) Test the hypothesis at the .05 level that the regression is zero.

(*d*) Test the hypothesis that the regression is linear.

(*e*) Interpret the above results, explaining what conclusions you have reached about high-school costs per pupil and size of high school.

3. A scatter diagram showing points plotted for pairs of measures in X_1 and X_2 are found to cluster closely about an inverted semicircle like a rainbow. The product-moment correlation coefficient is found to be zero.

(*a*) How do you account for the zero correlation?

(*b*) Does this mean that there is no correlation between the two variables?

(*c*) Which of the two correlation ratios would be useful in measuring correlation in this case?

(*d*) Why is the other one not useful?

4. Estimate the correlation between sex and intentions of going to college for high school A of Appendix G. Use three methods, the contingency coefficient, the phi coefficient, and the tetrachoric correlation coefficient. Explain why there are differences in results. How would you test the hypothesis that sex and intentions of going to college are independent on the basis of the high school A sample? Which of the three coefficients would be suitable for finding the correlation between college intentions and high school attended, using the data of Appendix G?

5. A test given to 200 students was found to have a standard deviation of 12. The 200 students were grouped on the basis of grades into two groups, "high academic standing" and "low academic standing." The 70 students in the high group had a mean score of 64 on the test. The mean score for the low group was 53. Why is it not logical to use the point-biserial coefficient in a case like this? Compute the biserial correlation between the test and academic standing.

6. Of 63 students taking a test, 32 answered the first item correctly and 31 incorrectly The data given below show the distributions of scores of the two groups on the total test. Does the first item measure the same thing that the total test measures? Compute the point-biserial coefficient of correlation and test the significance of the result at the .05 level.

Total Score	Item 1 Response	
	Correct	Incorrect
30–32	1	–
27–29	2	2
24–26	3	2
21–23	4	3
18–20	7	5
15–17	6	8
12–14	4	7
9–11	3	3
6– 8	2	1

7. Using the point-biserial correlation coefficient, find the correlation over all 159 subjects in Appendix G between intentions to go to college and general intelligence

(CTMM score). Test the hypothesis that this correlation is zero. To what population is this inference applicable?

8. Compute the contingency coefficient from the data of Ex. 2 of Chapter 10 to show the relationship between achievement in high school and occupation of father.

9. Compute the phi coefficient to show the correlation between Item *a* and Item *b* of a test from the responses shown below for 68 pupils.

	Item *a*	
Item *b*	Correct	Incorrect
Correct	24	6
Incorrect	18	20

10. Suppose that in the pilot training experiment of Section 8.1 the two examiners had been asked to rank the 10 students. Suppose, furthermore, that the rankings would correspond to the ranks of the grades in Table 8.1. Determine these ranks and find the rank correlation. Compare results with the product-moment coefficient. If the proper tables are available, test the hypothesis that the rank correlation is zero.

11. Drawings of 12 pupils were independently ranked by two judges. From the data below compute the correlation between the judges' rankings:

	A	B	C	D	E	F	G	H	I	J	K	L
First judge	1	4	7	2	8	6	12	3	9	5	10	11
Second judge	3	2	10	5	8	7	11	1	12	4	6	9

12. What is the correlation coefficient between the CTMM and the physical science test derived from all 159 observations in Appendix G? (See Ex. 2, Chapter 9.) Why would a simple average of the correlations of the four subgroups in Ex. 14 of Chapter 8 not be a good approximation of this? Suppose that only the r's and the n's of the four subgroups were given. How would you estimate r for the total group? Compute this estimate and compare with the result obtained directly from Appendix G.

REFERENCES

1. Cheshire, Leone, Milton Saffir, and L. L. Thurstone, *Computing Diagrams for the Tetrachoric Correlation Coefficient*, Chicago, University of Chicago Bookstore, 1933.
2. Dixon, Wilfrid J., and Frank J. Massey, *Introduction to Statistical Analysis*, New York, McGraw-Hill Book Co., 1951, pp. 260-261.
3. Guilford, Joy P., *Fundamental Statistics in Psychology and Education*, Second Ed., New York, McGraw-Hill Book Co., 1950, Chapter 13.
4. Olds, Edwin G., "Distribution of the Sums of Squares of Rank Differences for Small Numbers of Individuals," *Annals of Mathematical Statistics*, 9: 133–48, 1938.
5. Olds, Edwin G., "The 5% Significance Levels for Sums of Squares of Rank Differences and a Correction," *Annals of Mathematical Statistics*, 20: 117–118, March 1949.
6. Perry, Norman C., and William B. Michael, "The Reliability of a Point Biserial Coefficient of Correlation," *Psychometrika*, 19: 313-25, December 1954.
7. Walker, Helen M., and Joseph Lev, *Statistical Inference*, New York, Henry Holt and Co., 1953, Chapters 11 and 12.
8. Wert, James E., Charles O. Neidt, and J. Stanley Ahmann, *Statistical Methods in Educational and Psychological Research*, New York, Appleton-Century-Crofts, 1954, Chapters 14, 15, 16, and 17.
9. Yule, G. Udny, and Maurice G. Kendall, *An Introduction to the Theory of Statistics*, Thirteenth Ed., Revised, London, Charles Griffin and Co., 1947, Chapters 15 and 16.

CHAPTER 14

Partial and Multiple Correlation

Discussions of correlation to this point have concerned the relationships of only *two* variables. Such correlation is sometimes called *zero-order* correlation. This chapter considers some of the most commonly used elementary topics of *multivariate* analysis, that is, analysis concerned with three or more different measures for each individual in a population or in a sample.

Most of the principles of theory essential to understanding the topics of multiple regression, multiple correlation, and partial correlation which are discussed in this chapter are identical with those basic to the simpler correlation and regression problems discussed in Chapter 8. Thus a good way to prepare for the study of this chapter is to review systematically the underlying theory emphasized in Chapter 8.

14.1 THE THREE-VARIABLE PROBLEM

Suppose that for a group of objects or individuals we have three different measures, X_1, X_2, X_3, for instance, height, weight, and age of pupils. They might instead be scores on a reading test, an intelligence test, and a measure of study habits for a group of pupils, or assessed valuation per capita, income per capita, and retail sales per capita for a group of school districts.

We may be interested in the simple, or "zero-order," intercorrelation of the three measures. If so, we proceed by the methods of Chapter 8 and compute r_{12}, r_{13}, and r_{23}, the three intercorrelations among the three variables. However, we may suspect that the correlation between one pair of variables is influenced somehow by the relationship of each to the third. This is a problem of *partial correlation*. On the other hand, we might wish to know how well one measure can be predicted from both of the other two taken in combination. This problem is similar to the regression involving only two variables as in Chapter 8, except that now we have *two* independent variables instead of one. Moreover, instead of two possible regression equations, there would be three. There would be a regression for prediction of the first variable on the second and third, one for predicting the second on the first and third, and one for predicting the third on the first and second. This must be remembered in theory, but most often the nature of the regression problem is such that we are interested in only one of the three possible regression equations. In studies of local tax-paying ability in school finance, we are not interested in predicting a school district's personal income or volume of retail sales; we might, however, be interested in predicting a district's property valuation from other rationally defensible measures such as retail sales and personal income as a means of developing a measure free of the gross errors or inequities involved in the assessment of property. Hence, we usually label the first variable as the one in which interest centers, as the variable to be predicted, or as the *criterion* variable.

We will first study the problem of predicting one measure from two others. We will see how this may be generalized to multiple regression involving more than three variables.

14.2 MULTIPLE REGRESSION

In our discussion we will assume that we have particular interest in the first variable as a criterion and that we are interested in predicting it from the other two, at the same time recognizing that there are two other prediction problems which might be considered.

The prediction equation, or *multiple regression*, for predicting the first variable on the other two is

$$\hat{X}_1 = b_{12.3}X_2 + b_{13.2}X_3 + a \tag{14.1}$$

This follows the general idea of equation 8.9, the difference being that we have an additional term involving X_3 as an independent variable.

We also have two *regression coefficients*, $b_{12.3}$ and $b_{13.2}$. The subscripts differ from those used in Chapter 8 because the coefficients are different. In other words, the coefficient for the variable, X_2, in equation 14.1 ordinarily takes on a different numerical value from that of equation 8.14. The difference is due to the effect of the variable X_3. In the multiple-regression equation the coefficients are sometimes called *partial-regression coefficients* because each shows the regression of the dependent variable upon an independent variable with the effect of other variables eliminated. For instance, $b_{12.3}$ is the regression of variable *one* on variable *two* which would be found among many pairs of X_1, X_2, all having the same value of X_3. It represents the *net* regression of variable one on variable two.

This notation distinguishes the coefficients in equation 14.1 from those in the two possible three-variable regression equations. For instance, $b_{21.3}$ would be the coefficient of X_1 in the regression equation for predicting X_2 from X_1 and X_3.

To simplify the discussion we will consider all variables as deviations from their respective means. The above regression equation may then be written

$$\hat{x}_1 = b_{12.3}x_2 + b_{13.2}x_3 \tag{14.2}$$

This is the regression equation in deviation form for the three-variable problem and is comparable to equation 8.13. The *intercept* or the constant term, a, may be derived in a manner similar to that of equation 8.16 as follows:

$$a = \bar{X}_1 - b_{12.3}\bar{X}_2 - b_{13.2}\bar{X}_3 \tag{14.3}$$

The computational problem is to determine the partial regression coefficients. The theory involved is similar to that for the bivariate case which was represented in Chapter 8 graphically in two dimensions. In the present case, however, we must visualize a *three*-dimensional space, with three axes intersecting at right angles to one another at the location of the means of the three variables. We may think of the vertical axis as representing the scale for the measurement of X_1, another axis perpendicular to it for measuring X_2, and a third perpendicular to the plane defined by the other two axes, representing the measurement of X_3. Instead of a two-dimensional scatter diagram with a swarm of points in a plane, as we had in Chapter 8, we now have a swarm or cloud of points about the origin in a three-dimensional space. The location of each point in space is determined by three coordinates, the three variables measured as deviations from their means. The regression equation (14.2), in the three-dimensional representation, is a plane, whereas in the two-dimensional problem of Chapter 8 it was merely a straight line. Vertical distances from the plane to points in the three-dimensional space are *residuals*, that is, deviations from the regression equation, $X_1 - \hat{X}_1$, just as vertical distances from the regression line in the two-variable problem were residuals.

We assume that the distribution of points is trivariate normal. This means that the three bivariate distributions must all be normal and homoscedastic, the univariate distributions are normal, and the multiple-regression surface is a plane.

The mathematical theory of regression in the three-variable problem uses *least squares*, as it does in the simpler case. That is, computation of the regression coefficients is determined by the criterion that the sums of squares of residuals from the plane which they define is a minimum. The mathematical procedures in solving this problem are similar to that of Chapter 8. Details will not be presented here, but it should be understood that the procedures: (1) minimize sums of squares of residuals, and (2) maximize the correlation between predicted and actual values in the dependent variable, X_1.

The solution of the regression equation may be in gross score form or in deviation form, but there are advantages in expressing the regression equation in *standard score* form. If we express each variable in standard score form, substituting zs for x in equation 14.2 for both independent variables, then

$$\hat{z}_1 = \left(b_{12.3}\frac{s_2}{s_1}\right) z_2 + \left(b_{13.2}\frac{s_3}{s_1}\right) z_3$$

The coefficients of z_2 and z_3 (in parentheses) were altered by changing our measures to standard form. The symbols usually used for regression coefficients in standard form are[1]

$$\beta_{12.3} = b_{12.3}\frac{s_2}{s_1}$$

$$\beta_{13.2} = b_{13.2}\frac{s_3}{s_1}$$

Using this new notation for the partial regression coefficients in standard form, we obtain the equation

$$\hat{z}_1 = \beta_{12.3}z_2 + \beta_{13.2}z_3 \tag{14.4}$$

Similar to the mathematics for finding the least squares solution of Chapter 8, it can be shown that the coefficients in equation 14.4 must satisfy the so-called *normal* equations,

$$\begin{aligned}\beta_{12.3} + r_{23}\beta_{13.2} &= r_{12}\\ r_{23}\beta_{12.3} + \beta_{13.2} &= r_{13}\end{aligned} \tag{14.5}$$

In a numerical solution the values of r may be computed by the methods

[1] We depart here from the convention of using Greek letters for parameters and Roman letters for statistics. The Greek letter, beta, is so commonly used for regression weights or regression coefficients in standard form that we are adopting that practice even though it is inconsistent. It should be emphasized that these coefficients are derived from samples and hence are subject to sampling variation.

of Chapter 8 and the two *β coefficients* may be found by solving the two equations simultaneously. This is equivalent to the direct solution using

$$\beta_{12.3} = \frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2}$$

and

$$\beta_{13.2} = \frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2} \tag{14.6}$$

Computation for determining the multiple-regression equation in the three-variable problem is really quite simple, as the following example will demonstrate. For 321 students the means and standard deviations on three measures were as follows:

	$\bar{X}_i$	s_i
X_1 = grade-point average	18.51	11.24
X_2 = general intelligence test	100.62	15.83
X_3 = achievement test	24.22	6.15

We have, also, the intercorrelations of variables which we present as a *correlation matrix:*

	X_1	X_2	X_3
X_1	1.000	.583	.517
X_2	.583	1.000	.734
X_3	.517	.734	1.000

This matrix is "symmetrical," that is, corresponding figures in the upper right and lower left parts are equal, because $r_{ij} = r_{ji}$. Hence values below (or above) the diagonal are sometimes omitted for brevity.

Substituting in equation 14.6, we find the β coefficients as follows:

$$\beta_{12.3} = \frac{.583 - (.517)(.734)}{1 - .539} = .204/.461 = .443$$

and

$$\beta_{13.2} = \frac{.517 - (.583)(.734)}{1 - .539} = .089/.461 = .193$$

In standard score form the regression equation is, therefore,

$$\hat{z}_1 = .443z_2 + .193z_3$$

By definition of the β coefficients, they are readily converted to b coefficients from the relationships

$$b_{12.3} = \beta_{12.3}\frac{s_1}{s_2}$$

$$b_{13.2} = \beta_{13.2}\frac{s_1}{s_3} \tag{14.7}$$

Our partial regression coefficients are, hence,

$$b_{12.3} = (.443)(11.24)/(15.83) = .315,$$

and

$$b_{13.2} = (.193)(11.24)/(6.15) = .353$$

Substituting in equation 14.3, we find $a = -21.7$, and the regression equation is $\hat{X}_1 = .315X_2 + .353X_3 - 21.7$.

In using such a regression in the actual prediction of an individual's X_1 score, we would proceed as in Chapter 8. As an example, substituting in the above equation, an individual with a score on the general intelligence test of 112, and a score on the achievement test of 27, would have a predicted grade-point average of 23.1.

14.3 MULTIPLE CORRELATION

Our interest may now turn to the question of *how well* our regression equation predicts. This we may determine by correlating regression values, $\hat{X}_1$, and actual values X_1. We define the *multiple correlation coefficient*, $r_{1.23}$, as the correlation between the actual values of X_1 and the values predicted on the basis of the variables 2 and 3.

If individual scores were at hand, we could actually compute for each individual in our example the predicted value, or regression value, and correlate these with the actual values of X_1. The result would be the multiple-correlation coefficient. This would not be a very economical procedure, but it is good to know that the multiple-correlation coefficient is in fact a measure of the correlation of one variable with the best linear combination of other given variables.[1]

We may now draw upon principles emphasized in Sections 8.5 and 13.1. As with simpler correlation problems, the sum of squares of deviations (total) in the criterion or dependent variable is equal to the sum of squares of residuals plus the sum of squares of regression values. Thus

$$\Sigma x_1^2 = \Sigma x_{1.23\ldots m}^2 + \Sigma \hat{x}_1^2$$

where $x_{1.23\ldots m}^2$ is $(X_1 - \hat{X}_1)^2$, the square of the residual of X_1 from the regression value estimated from the remaining m measures. The other notation is as in Chapter 8.

[1] As with simple "zero-order" correlation between two variables only, relationships may be curvilinear. The study of multiple-curvilinear correlation is beyond the scope of this book. In three-variable curvilinear distributions, instead of a plane there would be a curved regression surface.

From this we may develop an equivalent definition of the multiple correlation coefficient on the principle underlying equation 13.2 for the *correlation index*. In this form we may write

$$r^2_{1.23\ldots m} = 1 - \frac{\Sigma x^2_{1.23\ldots m}}{\Sigma x^2_1} \tag{14.8}$$

The subscripts for multiple r indicate that it is the correlation of variable 1 with the remaining variables combined. This equation shows that, like the zero-order correlation, the closer the swarm of points are fitted by the *plane*, that is, the *smaller* the residuals, the *higher* the correlation. The three-variable geometric model must be generalized still further if there are more than three variables in the problem. In that case we are dealing with m-dimensional space and *hyper-planes*, but the general principles are the same.

Even equation 14.8 is not suitable for most computational purposes. It can be shown that the square of multiple r can be computed from

$$r^2_{1.23\ldots m} = \beta_2 r_{12} + \beta_3 r_{13} + \cdots + \beta_m r_{1m} \tag{14.9}$$

In a more complete notation we would designate β_2 as $\beta_{12.3\ldots m}$. The simpler though less precise notation will be used here, bearing in mind which variable is the dependent variable and which other variables are included as the independent variables.

Equation 14.9 gives the square of the multiple r as the sum of products of (a) correlations of independent measures with the dependent measure, and (b) the respective *standard* partial regression coefficients. This is also the ratio of *regression variance* to *total variance* in variable X_1. In terms of sums of squares

$$r^2_{1.23\ldots m} = \Sigma \hat{x}^2_1 / \Sigma x^2_1$$

Returning to our three-variable problem and using the abbreviated notation for our beta coefficients, we can find the multiple correlation coefficient by taking the square root of

$$\begin{aligned} r^2_{1.23} &= \beta_2 r_{12} + \beta_3 r_{13} \\ &= (.443)(.583) + (.193)(.517) \qquad (14.10) \\ &= .258 + .100 \\ &= .358 \end{aligned}$$

We see here that the proportion of "explained" variance in X_1 is .358.

Taking the square root, $r_{1.23} = .598$, which represents little higher predictive power than using X_2 alone, with correlation .583, or X_3 alone, with correlation .517. This reflects the high intercorrelation, .734, between the two predictors. The importance of this correlation appears clearly in the formula

$$r^2_{1.23} = \frac{r^2_{12} + r^2_{13} - 2r_{12}r_{13}r_{23}}{1 - r^2_{23}} \qquad (14.11)$$

which can be used in three-variable problems.

The third term in the numerator contains r_{23}, the correlation between independent measures, as a factor. The higher this correlation, the more it tends to reduce the multiple correlation, provided that both r_{12} and r_{13} are positive. Hence, in selecting batteries of tests or combinations of predictors, test constructors attempt to find measures which correlate *high* with the criterion but correlate *low* with one another.

14.4 THE STANDARD ERROR OF ESTIMATE

Researchers in education who work with large samples very often compute the standard error of estimate by formulas equivalent to taking the square root of the sum of squares of residuals divided by the number of cases in the sample,

$$s_{1.23\ldots m} = \sqrt{\Sigma x^2_{1.23\ldots m}/n}$$

An unbiased estimate is always to be preferred, differing only in that the residual sum of squares is divided by the appropriate number of degrees of freedom. The sum of squares of residuals may be computed easily from multiple r. From the development of equation 14.8, it is easily seen that

$$\Sigma x^2_{1.23\ldots m} = \Sigma x^2_1(1 - r^2_{1.23\ldots m}) \qquad (14.12)$$

and

$$\Sigma \hat{x}^2_1 = \Sigma x^2_1(r^2_{1.23\ldots m}) \qquad (14.13)$$

The appropriate number of degrees of freedom for equation 14.12 is $(n - k - 1)$, where k is the number of *independent* variables. In Section 8.6 with only one independent variable we saw that the number of degrees of freedom was $(n - 2)$. In the three-variable problem it is $(n - 3)$, and so on. Therefore, the square of the standard error of estimate is

$$s^2_{1.23\ldots m} = \frac{\Sigma x^2_1(1 - r^2_{1.23\ldots m})}{n - k - 1}$$

$$= \left(\frac{n - 1}{n - k - 1}\right) s^2_1(1 - r^2_{1.23\ldots m}) \qquad (14.14)$$

In the foregoing example we substitute in equation 14.14 to find $s_{1.23}^2 = (320/318)(126.34)(1 - .358) = 81.62$. Taking the square root of this, we find $s_{1.23} = 9.03$. With n as large as in this example, 321, the first factor in equation 14.14 has little effect, being 1.006. When n is large, the variance of estimate may thus be computed as $s_1^2(1 - r_{1.23\ldots m}^2)$, nearly.

14.5 TESTING THE SIGNIFICANCE OF MULTIPLE *r*

Although it is not necessary, as we shall see, to make an analysis of variance in order to test the significance of multiple r, that is a good approach for understanding the meaning of the appropriate test. We have already seen that we can partition the sum of squares of deviations in variable X_1 into the two components, (1) sum of squares due to regression, and (2) sum of squares due to residuals, and that these components can be derived from the total sum of squares and the square of the multiple-correlation coefficient.

The sum of squares of *residuals* could be computed as

$$\Sigma x_{1.23\ldots m}^2 = \Sigma X_1^2 - a\Sigma X_1 - b_2\Sigma X_1X_2 - \cdots - b_m\Sigma X_1X_m \quad (14.15)$$

The *total* sum of squares may be computed by methods already familiar to us, and by subtraction we find the sum of squares due to regression.

For the present we will use equations 14.12 and 14.13 to represent our components of sums of squares in a multiple-regression problem. Table 14.1 indicates the form of the analysis of variance.

TABLE 14.1

ANALYSIS OF VARIANCE FOR TESTING SIGNIFICANCE OF MULTIPLE CORRELATION

Source of Variation	Sum of Squares	Degrees of Freedom
Total	Σx_1^2	$n - 1$
Regression	$\Sigma \hat{x}_1^2 = r_{1.23\ldots m}^2 \Sigma x_1^2$	k
Residuals	$\Sigma x_{1.23\ldots m}^2 = (1 - r_{1.23\ldots m}^2)\Sigma x_1^2$	$n - k - 1$

The number of degrees of freedom is $(n - 1)$ for the *total* sum of squares. The degrees of freedom for *regression* is the number of *independent* measures, k, which is one less than the total number of measures in the problem, $(m - 1)$. This leaves $(n - k - 1)$ degrees of freedom for *residuals*, as we have already noted. In computing the sum of squares for residuals we lose a degree of freedom in computing the mean of X_1 and a degree of freedom for each of k regression coefficients (one for each of the k independent variables). We thus might have designated degrees of freedom for residuals as $(n - m)$.

From the information of Table 14.1 we could compute an estimate of a population variance from the *regression* sum of squares and another estimate from the *residual* sum of squares by dividing by the appropriate degrees of freedom. We could use the residual variance as error variance and make an F test of the significance of regression. This would test the hypothesis $H : \rho_{1.23\ldots m} = 0$, that there is no correlation whatever between X_1 and the other variables.

The actual computations can be simplified by canceling Σx_1^2 from numerator and denominator of the F ratio, leaving

$$F = \frac{r_{1.23\ldots m}^2/k}{(1 - r_{1.23\ldots m}^2)(n - k - 1)} = \frac{n - k - 1}{k}\,\frac{r_{1.23\ldots m}^2}{1 - r_{1.23\ldots m}^2} \tag{14.16}$$

where k is the number of degrees of freedom for the greater mean square, and $n - k - 1 = n - m$ degrees of freedom for error. Applying the test, we find that $F = \dfrac{.358/2}{.642/318} = 88.7$. For 2 and 318 d.f. we find that this is highly significant.

Although this gives us an over-all test of regression, it does not indicate to us the significance of individual regression coefficients. Computations for testing the significance of individual partial regression coefficients rapidly become cumbersome as the number of independent variables increases, but it is easy enough to test the significance of the two beta coefficients in a three-variable problem. The square of the standard error is the same for the two beta coefficients and is

$$s_\beta^2 = \frac{1 - r_{1.23}^2}{(1 - r_{23}^2)(n - 3)} \tag{14.17}$$

The appropriate test is the t test, where $t = \beta/s_\beta$ for $(n - 3)$ degrees of freedom. In our three-variable problem we substitute in equation 14.17 and take the square root to find $s_\beta = .0661$. For $\beta_{12.3}$ we find $t = (.443)/(.0661) = 6.7$. With 318 d.f. the t distribution is practically the same as the normal distribution. Therefore, we may look upon this ratio as a normal deviate, z, which is in excess of the critical value, 2.58 at the 1 percent level. Similarly, $\beta_{13.2}$ is found to have $t = 2.92$, which is significant at the 1 percent level. We therefore *reject* the two hypotheses that there is no partial regression of X_1 on the second variable and no partial regression of X_1 on the third variable.

14.6 SOLUTION FOR MORE THAN THREE VARIABLES

A general formula for a multiple-regression involving any number of variables, m, is

$$\hat{X}_1 = b_2X_2 + b_3X_3 + \cdots + b_mX_m + a \tag{14.18}$$

and this is only one of the m possible regression equations. We have used an abbreviated notation in our subscripts for the regression coefficients, and it is to be remembered that the coefficients depend upon which is the dependent variable. For instance, in a five-variable problem, b_2 could be either $b_{12.345}$ or $b_{32.145}$. In one case X_1 is the dependent variable, in the other X_3 is the dependent variable. It should be clear what is intended in equation 14.18.

Since the multiple-correlation coefficient is the measure of "accuracy of prediction," there are as many multiple-correlation coefficients as there are regressions. Hence, there are m different multiple-correlation coefficients, namely, $r_{1.23\cdots m}$, $r_{2.13\cdots m}$, $r_{3.12\cdots m}$, etc.

The general multiple-regression equation in *standard* form is

$$\hat{z}_1 = \beta_2z_2 + \beta_3z_3 + \cdots + \beta_mz_m \tag{14.19}$$

If the solution is made in terms of β coefficients, it is easy to convert to the b coefficients. The method is the same as in equation 14.7, that is,

$$b_j = \beta_j(s_1/s_j)$$

The constant term, a, in equation 14.18 is

$$a = \bar{X}_1 - b_2\bar{X}_2 - b_3\bar{X}_3 - \cdots - b_mX_m \tag{14.20}$$

A method of solution of regression coefficients is the solution of simultaneous equations, or *normal* equations, in terms of the β coefficients and the zero-order correlation coefficients. There are as many such equations as there are independent variables. The normal equations for a five-variable problem are as follows:

$$\begin{aligned}
\beta_2 + \beta_3r_{23} + \beta_4r_{24} + \beta_5r_{25} &= r_{12}\\
\beta_2r_{23} + \beta_3 + \beta_4r_{34} + \beta_5r_{35} &= r_{13}\\
\beta_2r_{24} + \beta_3r_{34} + \beta_4 + \beta_5r_{45} &= r_{14}\\
\beta_2r_{25} + \beta_3r_{35} + \beta_4r_{45} + \beta_5 &= r_{15}
\end{aligned}$$

An examination of the arrangement of terms in these equations will reveal the pattern which can be extended to any number of variables. For instance, for six variables the term $\beta_6 r_{26}$ would be added to the first equation, similar terms to the other equations, and we would add a fifth equation

$$\beta_2 r_{26} + \beta_3 r_{36} + \cdots = r_{16}$$

We illustrate a computation technique for finding the betas, known as the Doolittle solution, in Table 14.2. The problem is a prediction of X_1, grade-point average, from the same sample we have already used in our three-variable problem. We now add two additional variables with means and standard deviations as follows:

	$\bar{X}$	s
X_4, English usage	79.34	7.57
X_5, Music appreciation	32.13	4.21

The extended matrix of intercorrelation is now:

	X_1	X_2	X_3	X_4	X_5
X_1	1.000	.583	.517	.498	.315
X_2		1.000	.734	.852	.407
X_3			1.000	.744	.423
X_4				1.000	.435
X_5					1.000

The first step in the scheme presented in Table 14.2 is to enter on the first four lines the proper coefficient for the betas in the normal equation. For instance, in line X_2, the coefficient of the first term in the first normal equation is 1.0000. The coefficient of β_3 in the next term is r_{23}, which we find from our matrix is .734. The next term involves $r_{24} = .852$, and so on. In order to reduce the accumulation of rounding errors, we have added a zero to each entry, carrying the work to four decimal places. In the next-to-last column, headed $-X_1$, we enter the correlation of the dependent variable with each of the four independent variables in order—changing signs.

The first four lines, in short, are the normal equations, though the unknowns, the betas, do not appear explicitly in each term. Since the arrangement is symmetrical, it is unnecessary in the solution to record or use the coefficients in the lower left-hand portion of this arrangement, in italics.

The last column is a check column. It contains the algebraic sum of all the other entries in the same row.

TABLE 14.2

Doolittle Solution of Regression Coefficients

Line	Procedure	X_2	X_3	X_4	X_5	$-X_1$	Check
X_2	(Enter r_{ij} in scheme of normal equations.)	1.0000	.7340	.8520	.4070	−.5830	2.4100
X_3		*.7340*	1.0000	.7440	.4230	−.5170	2.3840
X_4		*.8520*	*.7440*	1.0000	.4350	−.4980	2.5330
X_5		*.4070*	*.4230*	*.4350*	1.0000	−.3150	1.9500
I (*a*)	Line X_2	1.0000	.7340	.8520	.4070	−.5830	2.4100
(*b*)	Line (*a*) × (−1.0000)	−1.0000	−.7340	−.8520	−.4070	.5830	−2.4100
II (*c*)	Line X_3		1.0000	.7440	.4230	−.5170	2.3840
(*d*)	Line (*a*) × (−.7340)		−.5388	−.6254	−.2987	.4279	−1.7689
(*e*)	Lines (*c*) + (*d*)		.4612	.1186	.1243	−.0891	.6151*
(*f*)	Line (*e*) × (−2.1683)		−1.0000	−.2572	−.2695	.1932	−1.3337*
III (*g*)	Line X_4			1.0000	.4350	−.4980	2.5330
(*h*)	Line (*a*) × (−.8520)			−.7259	−.3468	.4967	−2.0533
(*i*)	Line (*e*) × (−.2572)			−.0305	−.0320	.0229	−.1582
(*j*)	Lines (*g*) + (*h*) + (*i*)			.2436	.0562	.0216	.3215*
(*k*)	Line (*j*) × (−4.1051)			−1.0000	−.2307	−.0887	−1.3198*
IV (*l*)	Line X_5				1.0000	−.3150	1.9500
(*m*)	Line (*a*) × (—.4070)				−.1656	.2373	−.9809
(*n*)	Line (*e*) × (−.2695)				−.0335	.0240	−.1658
(*o*)	Line (*j*) × (−.2307)				−.0130	−.0050	−.0742
(*p*)	Lines (*k*)+(*m*)+(*n*)+(*o*)				.7879	−.0587	.7291*
(*q*)	Line (*p*) × (−1.2692)				−1.0000	.0745	−.9254*
	Back solution						
(*r*)	From line (*q*)	$\beta_5 =$				=	.0745
(*s*)	Substitute in line (*k*)	$\beta_4 = (.0745)(-.2307) + (-.0887)$				=	−.1059
(*t*)	Substitute in line (*f*)	$\beta_3 = (-.1059)(-.2572) + (.0745)(-.2695) + (.1932)$				=	.2003
(*u*)	Substitute in line (*b*)	$\beta_2 = (.2003)(-.7340) + (-.1059)(-.8520) + (.0745)(-.4070) + .5830$				=	.4959

$$(.4959)(1.0000) + (.2003)(.7340) + (-.1059)(.8520) + (.0745)(.4070) = .5830^*$$
$$(.4959)(.7340) + (.2003)(1.0000) + (-.1059)(.7440) + (.0745)(.4230) = .5170^*$$
$$(.4959)(.8520) + (.2003)(.7440) + (-.1059)(1.0000) + (.0745)(.4350) = .4980^*$$
$$(.4959)(.4070) + (.2003)(.4230) + (-.1059)(.4350) + (.0745)(1.0000) = .3150^*$$

* Computation check.

The computation follows a cyclical pattern of operations. Horizontal lines separating groups of rows (*a*) through (*q*) into the cycles. The pattern of computation may be learned through following computations in Table 14.2, and extending to problems involving more variables.

The next step in the procedure involves recopying line X_2, in line (*a*). We then enter in line (*b*) the values in line (*a*) with each sign changed. This completes the first cycle.

We now proceed to the second cycle, copying all entries (except the italicized ones) of line X_3, in line (*c*). This leaves blank the column headed X_2. Our work will thus start in this cycle with entries in column X_3, and proceed through each column to the right.

We find in the first column for this cycle (column X_3) that the *last figure in the previous cycle* is —.7340. Our next step is to multiply line (*a*) by this, entering results in line (*d*), except for column X_2. Line (*e*) is the algebraic sum, column-for-column, of entries in lines (*c*) and (*d*).

At this point our check column is examined to see if we have made computational errors. The check figure in line (*e*), beside being the sum of the other figures in the row, is also the sum of the two figures in the check column in lines (*c*) and (*d*), within an allowable error for rounding. The asterisk indicates a check. Permissible discrepancies due to rounding are in the neighborhood of .0005. Likewise, summing algebraically across line (*f*) we get an acceptable check with the figure in the last column in line (*f*) computed by the procedure given in the procedure column at the left of the table.

The final step in cycle II is to divide line (*e*) by the first entry in the row, .4612, with its sign changed. Results are entered in line (*f*). This is the same as multiplying through by the negative of the reciprocal of .4612, which is —1/.4612 = —2.1683. This completes cycle II.

In cycle III we continue by first copying down from line X_4 the next unused row of coefficients, entering these in line (*g*), and as before omitting italicized entries. We now begin computation in column X_4 as there are no entries in columns X_2 and X_3 in this cycle. Our reference column for computing in cycle III is thus column X_4. Looking above in this column to the first cycle, we find the last entry to be —.8520. We multiply the entries in line (*a*) by this and enter the results in line (*h*). Similarly, in this same column, we find the last entry in cycle II to be —.2572. We multiply through by this figure all entries in the line next above it and to the right, entering the results in line (*i*). We now sum in each column the entries in the first three lines of cycle III to obtain line (*j*). We then multiply line (*j*) by the negative reciprocal of the first entry, .2436, to complete entries for this cycle in line (*k*).

In cycle IV we proceed in a similar manner, first copying entries from line X_5, beginning in column X_5. Working in this column in the previous cycles, we find that we multiply line (*a*) by $-.4070$, line (*e*) by $-.2695$, and line (*j*) by $-.2307$. After these results are entered, we sum each column and enter the sums in line (*p*). The last line is found by multiplying line (*p*) through by the negative reciprocal of .7879. As we go along, we check in the last two lines in each cycle.

This completes what is called the *forward solution*. The result of this work gives us β_5, which we read in the $-X_1$ column in line (*q*) to be .0745. We are now ready for the *back solution*, which will give us the rest of the beta coefficients. In line (*r*) we enter β_5, which we have already found. For line (*s*) we go to line (*k*), multiply the entry in column X_5 by β_5, and add the entry in column $-X_1$. This gives us β_4. For line (*t*) we go to the last line of the next previous cycle, line (*f*) of cycle II, multiply the entry in column X_4 by β_4 add the product of the entry in column X_5 and β_5, and add to that the entry in column $-X_1$. The result is β_3. Working in line (*b*) in the same manner we find β_2.

At the bottom of Table 14.2 is a final check of computation. It is simply a substitution of the given r's and the derived β's in the four normal equations to see that they check.

We compute the b coefficients by multiplying each beta coefficient by the proper ratio of standard errors. For instance, $b_2 = \beta_2(s_1/s_2) = (.4959)(11.24)/(15.83) = .352$. In a similar manner we find $b_3 = .366$, $b_4 = -.157$, and $b_5 = .199$. By means of equation 14.20 we substitute to find $a = -19.70$. The complete regression is, therefore, $\hat{X}_1 = .352X_2 + .366X_3 - .157X_4 + .199X_5 - 19.70$. When this method of computation is used, a convenient way of finding the multiple-correlation coefficient is given by equation 14.9. We find $r^2_{1.2345} = .289 + .104 - .053 + .023 = .363$. The square root of this is $.602 = r_{1.2345}$.

It is of interest to note that the two new variables, X_4 and X_5, add little to what was available in X_2 and X_3. They reveal some of the elements measured in the other two variables. Their β weights are relatively low compared to the other two. By adding them, the multiple correlation is increased only from .598 to .603.

14.7 PARTIAL CORRELATION

In Section 8.8 it was emphasized that the magnitude of a product-moment correlation coefficient depends upon the nature of the population concerned, particularly with reference to variables other than the two correlated. The correlation between an intelligence test and a reading test would be expected to be much higher if the group of subjects ranged widely in chronological age and grade level than if they were all of about the same age. In other words, a high correlation between the two tests may be due partly to the correlation of each of them with the third variable, chronological age. This creates a difficulty of interpretation which can sometimes be solved by the experimenter through careful definition of his research objective so that he selects samples in a way to "control" extraneous factors, but this is not always possible.

Contamination arising from correlation with a third or several variables is a common possibility in educational research. In a physical education study, for example, we might find a sizable correlation between a grip test and height. On further examination of the records from our group of subjects, we would find even a higher correlation of the grip test with weight, and a high intercorrelation between height and weight of the subjects. The correlation of our measure of *strength* with height might have disappeared if all subjects had been of the same weight.

Another example is the correlation of an appraisal of educational programs with current expense per pupil in a group of school systems. In one study, such a correlation was found to be .587.[1] However, this correlation was obtained for a group of school districts very heterogeneous as to size and wealth. When a correlation coefficient was derived "within groups of school systems comparable as to size and wealth," the correlation was found to be only .175. In other words, some of the common variance exhibited in the correlation between school program and expenditure was explained by other factors.

If we have three variables, X_1, X_2, and X_3, and if we are interested in finding out the correlation between X_1 and X_2, *with the effects of* X_3 *eliminated*, we can use regression methods. We could find the regression of X_1 on X_3. The residuals from this regression would represent variations in X_1 *independent of*, or *not predictable by* X_3. Residuals, as we have seen in our previous study of correlation, represent in a statistical way that part of measures which cannot be explained by the independent variable. Similarly, we could find a regression of X_2 on X_3 and compute for our sample the residuals from this line. The results would be two sets of residuals, $x_{1.3}$ and $x_{2.3}$, each of which would represent those parts of X_1 and X_2, respectively, not related to the third variable, X_3. The correlation of these residuals would be a *partial-correlation coefficient*. As is sometimes said, it would show the correlation between X_1 and X_2 with the effects of X_3 removed.

The procedure might be even more general for we could use multiple-regression methods to eliminate "effects" of several variables. For

[1] Paul R. Mort and Francis G. Cornell, *American Schools in Transition*, New York, Teachers College, Columbia University, 1941. 465 pp.

instance, we might derive a coefficient which would be the correlation of the two residuals, $x_{1.345}$ and $x_{2.345}$, to find the "net" or partial correlation between variables X_1 and X_2 with the effects of the other three measures eliminated.

If only a few variables are involved, the partial-correlation coefficient is quite easily computed. With three variables, the partial-correlation coefficient for X_1 and X_2 with the effect of X_3 eliminated is

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}} \tag{14.21}$$

A careful distinction should be made between the notation used here and the notation used for multiple correlation, where $r_{1.23}$ represents a correlation of the first variable with both the others. Equation 14.21 gives *first-order* partial correlation (first order because there is only one variable eliminated) in terms of the *zero-order* correlations, the simple correlations between the variables.

A careful examination of equation 14.21 reveals a wide variety of changes that can be made from the zero-order correlation, r_{12}, to the partial correlation, $r_{12.3}$. For instance, if r_{12} is positive, but not great, and if r_{13} likewise is positive, but small, but if r_{23} is sufficiently high, $r_{12.3}$ may be negative. On the other hand, if either r_{13} or r_{23} is zero or very near zero, little change may result from eliminating the third variable. It is good to remember that the partial-correlation coefficient can be either greater or less than the zero-order correlation coefficient.

Some examples will illustrate the types of variations which may be expected from partial-correlation analysis. The first example is from a study of achievement, that is, the "proficiency" of military personnel trained as radio and radar maintenance mechanics. Among several tests administered to a group of mechanics who had completed training were the following: (1) a performance-like problem-solving test, (2) a test of understanding of checking and testing procedures which mechanics used in trouble-shooting, and (3) a basic knowledge or "information" type test. The correlation between the first two measures, r_{12}, was found to be .514, but the other two coefficients were even higher: $r_{13} = .525$, $r_{23} = .632$. Looking upon $r_{12} = .514$ as the square root of a variance ratio, we suspect that some of the relationship between the "common variance" of the first two measures is explained by the third. Substituting in equation 14.21, we find $r_{12.3} = .28$. Part of the observed correlation between X_1 and X_2 was explained by a common factor reflected in the third measure, the basic knowledge test. Eliminating the effect of the third measure, the correlation is considerably reduced.

In a study of human relations in the administration of school systems, there were three measures of teacher participation in decision making: (1) a measure of the "influence," that is, the degree of effect which teacher activity had upon final policy decision or action concerning various parts of the school program, (2) a measure of "degree of participation," that is, the amount of time and energy teachers devoted to decision making concerning school policy, and (3) the amount of "responsibility" delegated to individual teachers for decision making. Such measures were obtained for a sample of teachers in several school systems. The correlation between X_1, effectiveness, and X_2, degree of participation, was found to be .52. The other zero-order correlations were $r_{13} = .23$ and $r_{23} = .40$. The speculation was that the correlation between amount of influence and amount of activity (.52) might have been explained in part by variation in amount of responsibility or authority delegated to individual teachers. Substituting in equation 14.21, we find $r_{12.3} = .48$. Though there is some reduction in the correlation when the effect of the third variable is removed, the result is essentially unchanged.

In a study of Krathwohl,[1] a group of 308 subjects was measured on (1) English achievement, (2) vocabulary, and (3) indexes of industriousness in English. The intercorrelations were found to be $r_{12} = .58$, $r_{13} = .06$, and $r_{23} = -.47$. For 83 of the subjects who were classified as "industrious," the correlation between English achievement and vocabulary was .77. Likewise, a middle group or "normal" group in industriousness comprising 156 of the subjects showed $r_{12} = .66$, whereas the 69 "indolent" subjects showed a correlation between the first two variables of .56. This suggests that where there is "comparability as to industriousness" the correlation between vocabulary and English achievement is higher than the correlation of .58 observed for the total group. Substituting in equation 14.21, we find $r_{12.3} = .78$. This lends support to the notion that the correlation between English achievement (X_1) and vocabulary (X_2) was depressed by variations in "industriousness."

In this study, multiple-correlation coefficients were also reported. For instance, for all 308 cases $r_{1.23} = .70$. The reader should be able to distinguish between this and the partial correlation coefficient. Substituting in equation 14.11, we find the multiple r to be .69, differing from that reported because of rounding errors.

By changing the order of subscripts (equation 14.21) we may find in the foregoing three-variable examples the two other partial-correlation coefficients, $r_{13.2}$ and $r_{23.1}$. The order of designation of variables is,

[1] William C. Krathwohl, "Relative Contributions of Vocabulary and An Index of Industriousness for English to Achievement in English," *The Journal of Educational Psychology*, 42: 97-104, February 1951.

of course, arbitrary and at the investigator's convenience. Partial correlations of higher order may be computed in a similar manner. For instance,

$$r_{12.34} = \frac{r_{12.4} - r_{13.4}r_{23.4}}{\sqrt{1 - r_{13.4}^2}\sqrt{1 - r_{23.4}^2} \tag{14.22}$$

The similarity between the *partial-correlation* coefficient and the *partial-regression* coefficient is no accident. Each is a function of the other and can be computed from the other. A comparison of equation 14.6 and 14.21 shows the close relationship between the β coefficient and the partial r. Both have similar interpretations in that they represent the amount of variance common to two primary variables with the effects of the secondary variables eliminated.

14.8 TESTING THE SIGNIFICANCE OF THE PARTIAL-CORRELATION COEFFICIENT

Significance tests with partial-correlation coefficients are similar to those for zero-order correlation coefficients. However, it must be remembered that the number of degrees of freedom for partial correlation is less than that for zero-order correlation from the same sample. The number of degrees of freedom in partial correlation is reduced by the number of variables "partialled out." In Chapter 9, we saw that the number of degrees of freedom for testing the significance of the zero-order correlation coefficient, that is, only 2 variables, is $(n - 2)$. The same test may be made with partial-correlation coefficients with d.f. $= (n - m)$, where m equals the number of variables involved. The t test given in Section 9.4 would thus be for partial r

$$t = r_{12.34\ldots m}\frac{\sqrt{n - m}}{\sqrt{1 - r_{12.34\ldots m}^2}}; \quad \text{d.f.} = (n - m) \tag{14.23}$$

Fisher's z' transformation may be used with partial-correlation coefficients. The standard error of the z' is $1/\sqrt{n - m - 1}$.

This section is the last of the statistical theory treated in this book. It completes the coverage of topics which are essential for a working knowledge of educational statistics and essential for additional study of statistics. It completes neither the study of statistics in general nor the study of multivariate analysis. The reader who has progressed satisfactorily to this point should be able to move easily on to more advanced topics.

He will find *factor analysis* important to study if he is interested in tests and measurements or if he has problems of digesting the dimensionality of several "independent" variables, and references 1 and 2 good introductions to this subject. He will find it interesting to know about the *discriminant function* and *discriminant analysis* (reference 3). This knowledge will come in handy when he has a "qualitative" dependent variable and several "quantitative" independent variables. Discriminant analysis will be found to be an extension of the ideas of this chapter and Chapter 12. He will want to know about the *analysis of covariance* if he needs to account for one or more variables not otherwise "controlled" in experimental design.

These and several other topics of interest in education he will find in references at the end of chapters, or in current issues of the *Journal of the American Statistical Association, The Journal of Experimental Education, Psychometrika, and The Review of Educational Research.*

Finally, if he intends to specialize in educational statistics, he will wish to increase his understanding of the subject through the study of *mathematical* statistics.

EXERCISES

1. The means and standard deviations of a sample of 500 vocational students on (1) scores on a performance test in the operation of a machine, (2) general intelligence test scores, and (3) scores on a functional knowledge test were as follows:

$$\bar{X}_1 = 24.6 \qquad \bar{X}_2 = 82.4 \qquad \bar{X}_3 = 52.8$$

$$s_1 = 8.3 \qquad s_2 = 12.6 \qquad s_3 = 16.2$$

The intercorrelations were $r_{12} = .546$; $r_{13} = .671$; $r_{23} = .368$.

(*a*) Find the regression equation for predicting performance test scores.

(*b*) Compute the multiple-correlation coefficient.

(*c*) Test the hypothesis that the population multiple correlation is zero.

(*d*) Compute the three partial-correlation coefficients. Are they statistically significant?

2. The following body measurements were taken for 300 twenty-year-old male students entering a Midwest university over a five-year period: X_1, weight; X_2, height; X_3, chest (inspiration); and X_4, waist. The means, standard deviations, and intercorrelations of these measures based on the 300 students were as shown in the table.

	X_2	X_3	X_4	$\bar{X}$	s
X_1	.459	.816	.859	156.60	24.52
X_2		.339	.276	69.67	2.34
X_3			.747	38.17	2.56
X_4				30.41	2.93

(*a*) Find the regression equation for predicting weight from the other three measures.

(*b*) How accurately does the regression equation in *a* predict weight?

(*c*) Which of the three independent variables contributes most in predicting weight?

(*d*) Drop the least important independent variable and find the regression equation for predicting weight on the remaining two measures.

(*e*) Study your previous results to this exercise and see if you can estimate which partial-correlation coefficient is highest, $r_{13.24}$ or $r_{14.23}$?

(*f*) Test the hypothesis that the multiple correlation between weight and the other three variables is zero.

3. In what respects is the multiple-correlation coefficient similar to the: (*a*) Zero-order correlation coefficient? (*b*) Correlation index? (*c*) Correlation ratio?

4. Can the multiple-correlation coefficient be negative? Why?

5. What is the advantage in using beta coefficients in multiple-regression problems?

6. From a sample of 150 individuals, a multiple correlation with five independent measures is found to be .85. If the variance of the dependent variable is 18, what is the standard error of estimate?

7. Investigator A found the correlation between two variables to be $r_{12} = 0$. Investigator B, repeating the study, included a third measure and found $r_{12} = 0$; $r_{13} = -.43$; $r_{23} = .54$. How would you answer the question as to whether or not there is a correlation between X_1 and X_2? (Assume the sampling error to be negligible.)

REFERENCES

1. Fruchter, Benjamin, *Introduction to Factor Analysis*, New York, D. Van Nostrand Co., 1954, Chapters 1 and 4.
2. Guilford, Joy P., *Psychometric Methods*, Second Ed., New York, McGraw-Hill Book Co., 1954, Chapter 16.
3. Johnson, Palmer O., *Statistical Methods in Research*, New York, Prentice-Hall, 1949, Chapter 14.
4. Johnson, Palmer O., and Robert W. B. Jackson, *Introduction to Statistical Methods*, New York, Prentice-Hall, 1953, Chapter 13.
5. Walker, Helen M., and Joseph Lev, *Statistical Inference*, New York, Henry Holt and Co., 1953, Chapter 13.
6. Yule, G. Udny, and Maurice G. Kendall, *An Introduction to the Theory of Statistics*, Thirteenth Ed., Revised, London, Charles Griffin and Co., 1947, Chapter 14.

CHAPTER 15

Collecting and Reporting Statistical Data

The education profession utilizes many methods of assembling original data. Subjects are tested directly by group or individual tests, information is assembled from direct observation, and the interview is a common source of information. The commonest device seems to be the paper and pencil report form, test, or questionnaire. Abuses of the questionnaire have been listed in educational literature for many years. However, some type of prepared form or questionnaire will continue to be important in many types of operating or research work in education involving statistics.

15.1 PLANNING THE COLLECTION OF DATA

There is no reason why some of the chief ideas regarding good planning of collection instruments cannot become more widely understood among members of the profession. Too frequently "questionnaire" or "status" studies in education are undertaken simply because the investigator has not clearly defined his problem and thought through the question of what data specifically he needs for solving it. It is certainly important to have information about the teaching of mathematics in all the high schools in a state. but a specialist in the teaching of mathematics or some other subject is usually running out of worthwhile, testable hypotheses in his field when he resorts to a collection of data regarding *what* is taught and *how* it is taught.

On the other hand, in state agencies and the U.S. Office of Education there are important functions for the collection of basic descriptive statistics regarding schools. These data, like the federal census, provide basic quantitative information for the use of research workers and persons in education responsible for the development and administration of educational programs.

The Federal Government has a system of review of all forms and questionnaires, required by an act of Congress, the "Federal Reports Act" of 1942. It came about during the early parts of World War II chiefly as a result of the complaints of business and industrial establishments of the many government forms and questionnaires. Such organizations found, because of lack of coordination of federal agencies, an overwhelming demand for extensive and overlapping information. As a means of administering requirements of this Act, there is a Division of Statistical Standards in the Bureau of the Budget with the responsibility of approving and coordinating reporting plans for all federal agencies. Fortunately, the Division has been staffed with professional people whose point of view has leaned toward constructively improving the federal information-collecting services in addition to exercising their statutory function of controlling reports.

In the U.S. Office of Education, criteria for the evaluation of report forms and questionnaires were widely publicized among the professional staff in an attempt to develop an appreciation for the goal of improving forms and reports. An adaptation of these criteria is listed here because they cover most of the cautions which beginning statistical workers in education should have before them..

In one major respect, the collection device is of statistical importance. Statistical method, viewed as a science, enters into the design of experiments and investigations from the very start. The most effective uses of statistical method are those in which statistical theory influences the choices of (1) the data to be collected, (2) the populations from which they are to be assembled, and (3) the methods by which they are to be summarized and analyzed. Statistical method and the theory or subject matter of the field under investigation interact in the planning and execution of an efficient statistical study. In short, the time for statistical ideas in the solving of a problem is at the initial planning. Too often, even the most expert statistical consultant is unable to salvage data after the survey has been completed on a design wanting in some of the essential statistical ideas.

The forms and schedules used in a study are merely the means by which isolated facts are obtained. As such, they can be only as good as the program of research or reporting of which they are a part. It works the other way too, for no study or survey can be any better than the schedules of information upon which it is based.

The following list of principles of criteria for schedules was prepared after a review of the difficulties which had been encountered in the clearance of forms and a study of the most common weaknesses in the preparation of forms.

A. Function

Attention should be given to the purposes and general need of a form, and its use should be interpreted in terms of the intended end product.

1. *Program:* Is the administrative program or research plan specific enough so that the need for the form may be viewed in complete context of an operating plan?

2. *Importance:* Has the significance of information to be obtained from the form been judged with reference to *importance* to the program or plan?

3. *Relationship to objectives:* Are the data to be reported definitely related to objectives of the program or plan?

4. *Population:* Has attention been given to sampling in preference to complete coverage, in delineating the population to be covered?

5. *Research:* Has the form, even if developed primarily for administrative purposes, been reviewed in terms of its use as a source of statistical data for research?

B. Development

Organizations in which forms are developed in volume should follow a methodical system of procedure for form development. Persons working independently should utilize specialists who may be of assistance to them.

1. *Organization:* Is there a definite understanding among members of the staff regarding responsibilities with reference to form preparation and an organizational pattern through which forms are developed?

2. *Operating staff:* Has the form been cleared with persons who are to use the form, even though the form was not initiated by them?

3. *Technical staff:* Has the form been cleared by specially trained personnel? (Usually this staff would be in a research or statistical division or specially trained persons on the campus of a higher educational institution.)

4. *Respondent groups:* Have superintendents, principals, teachers or others who fill out the form been consulted concerning adequacy of the form and the procedures involved?

5. *Practical considerations:* Is the form too intricate to be effective? Does the form contain any irrelevant details which are unessential to final results?

6. *Experimentation:* Has a preliminary draft been tested if experience with a similar type of form cannot be drawn upon?

C. Plan of Use

A form or reporting procedure should be considered in the light of a specific plan for its use and the methods by which the plan will be put into effect.

1. *Tabulation plan:* Has a presurvey tabulation plan been drawn up and has the form been checked against it?

2. *Availability of data:* Has it been ascertained that information requested in the form cannot be obtained by use of existing records or available unassembled data?

3. *Facilities for tabulation:* Is there sufficient personnel and mechanical equipment for processing and tabulating data requested in the form?

4. *Item essentiality:* Can every item of the form be defended in terms of proposed actual use?

5. *Cost:* Has the total cost of obtaining the information (including man-hours of personnel already on payroll or in schools and school systems) been weighed, even if only crudely, against the value of information to be obtained?

6. *Motivation of respondents:* Has consideration been given to the problem of securing sufficient interest to insure cooperation of those who will fill out the form?

7. *Distribution and return:* Does the plan of use provide an efficient system of distribution and return of information?

8. *Coverage control:* Is there an efficient system of information planned for keeping track of returns?

D. Design

Forms and reporting procedures should be reviewed in terms of the total operating program or field of information which they are to serve.

1. *Structure:* Does the form dovetail into the existing structure of the reporting systems of schools, colleges, or school systems involved?

2. *Consistency:* Is each new form consistent with others in existence or previously used, with reference to: (*a*) administrative requirements, (*b*) the research design, (*c*) definitions and terminology, and (*d*) reporting time schedules?

3. *Utility:* Does the format facilitate filling in and handling (coding, filing, tabulating, transcribing, etc.)?

4. *Esthetic balance:* Is composition and lay-out designed for psychological appeal?

5. *Simplicity:* Is terminology understandable and simple and is form easy to fil! out?

E. Mechanical Suggestions

In applying the above suggestions, it is necessary to consider certain specific details of form construction.

1. *Objectivity:* As far as possible, restrict forms to readily determinable factual information. Avoid ambiguous items and those which are likely to invite error in reporting. If opinions are wanted, this should be clearly recognized and clearly indicated, and limitations of the questionnaire approach duly considered.

2. *Definition of items:* Units of description or measurement for gathering quantitative facts must be carefully defined. It is helpful to define items operationally, that is, in terms of the method by which data are to be obtained.

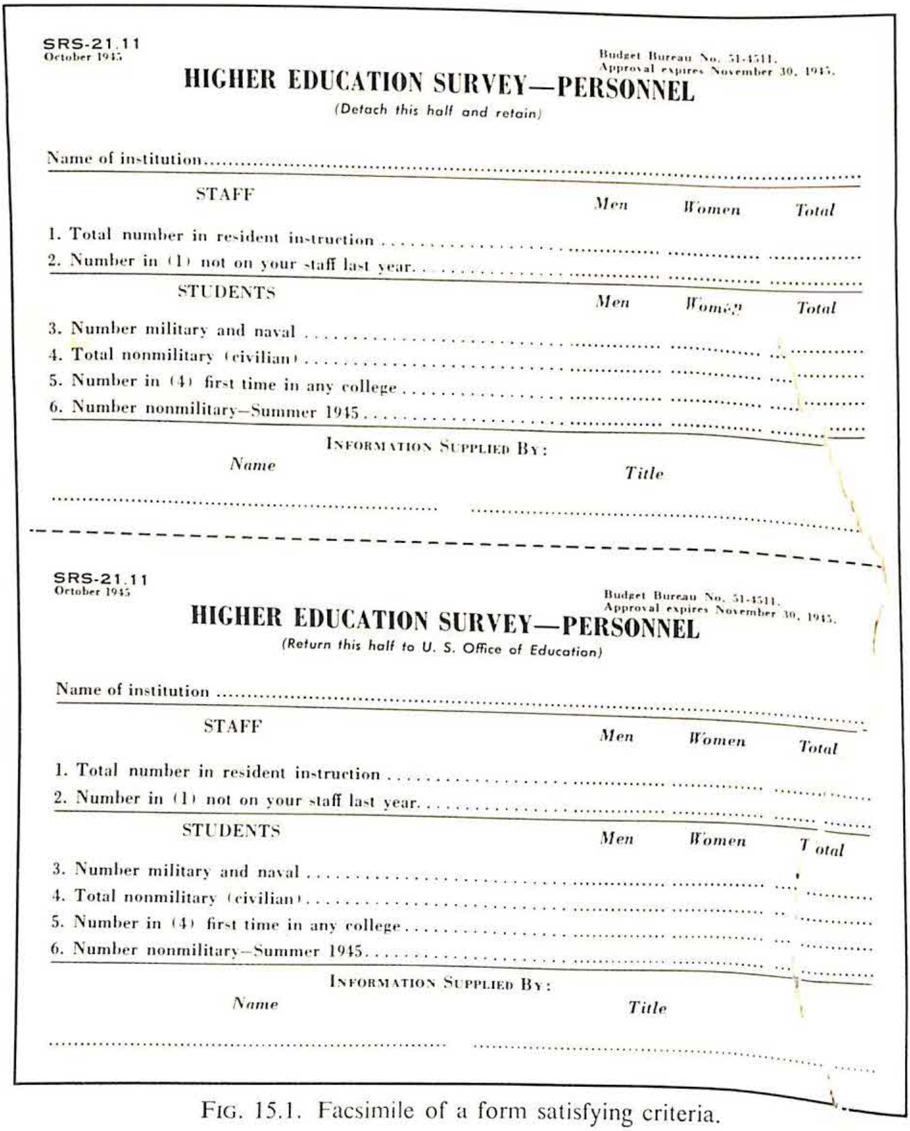
3. *Statistical suitability:* Items should be direct measures in preference to such derived measures as percentages and averages. Items must be defined so as to contribute best to the plan of statistical treatment.

4. *Computations:* Avoid requiring respondents to do the computing. For example, ask for (*a*) current expense, and (*b*) number of pupils in average daily attendance, but have (*c*) expenditures per pupil computed by staff tabulating results.

5. *Instructions:* Complete, simple, and understandable instructions are essential. That some respondents will not read them is no argument against them because of the confusion saved those who will.

6. *Identification:* Forms should be explicitly identifiable with adequate titles, identification numbers, sponsoring agency or authority, information showing by whom to be used, to whom to return, and when to be used, etc.

7. *Authentication:* Administrative documents should provide for authentication by signature and/or jurat of responsible authority. Space should be provided for signature of respondent, but in appropriate cases permission may be given to omit signature.



SRS-21.11
October 1945

Budget Bureau No. 51-4511.
Approval expires November 30, 1945.

HIGHER EDUCATION SURVEY—PERSONNEL

(Detach this half and retain)

Name of institution..

STAFF	*Men*	*Women*	*Total*
1. Total number in resident instruction			
2. Number in (1) not on your staff last year			

STUDENTS	*Men*	*Women*	*Total*
3. Number military and naval			
4. Total nonmilitary (civilian)			
5. Number in (4) first time in any college			
6. Number nonmilitary—Summer 1945			

INFORMATION SUPPLIED BY:

Name .... *Title* ....

SRS-21.11
October 1945

Budget Bureau No. 51-4511.
Approval expires November 30, 1945.

HIGHER EDUCATION SURVEY—PERSONNEL

(Return this half to U. S. Office of Education)

Name of institution..

STAFF	*Men*	*Women*	*Total*
1. Total number in resident instruction			
2. Number in (1) not on your staff last year			

STUDENTS	*Men*	*Women*	*Total*
3. Number military and naval			
4. Total nonmilitary (civilian)			
5. Number in (4) first time in any college			
6. Number nonmilitary—Summer 1945			

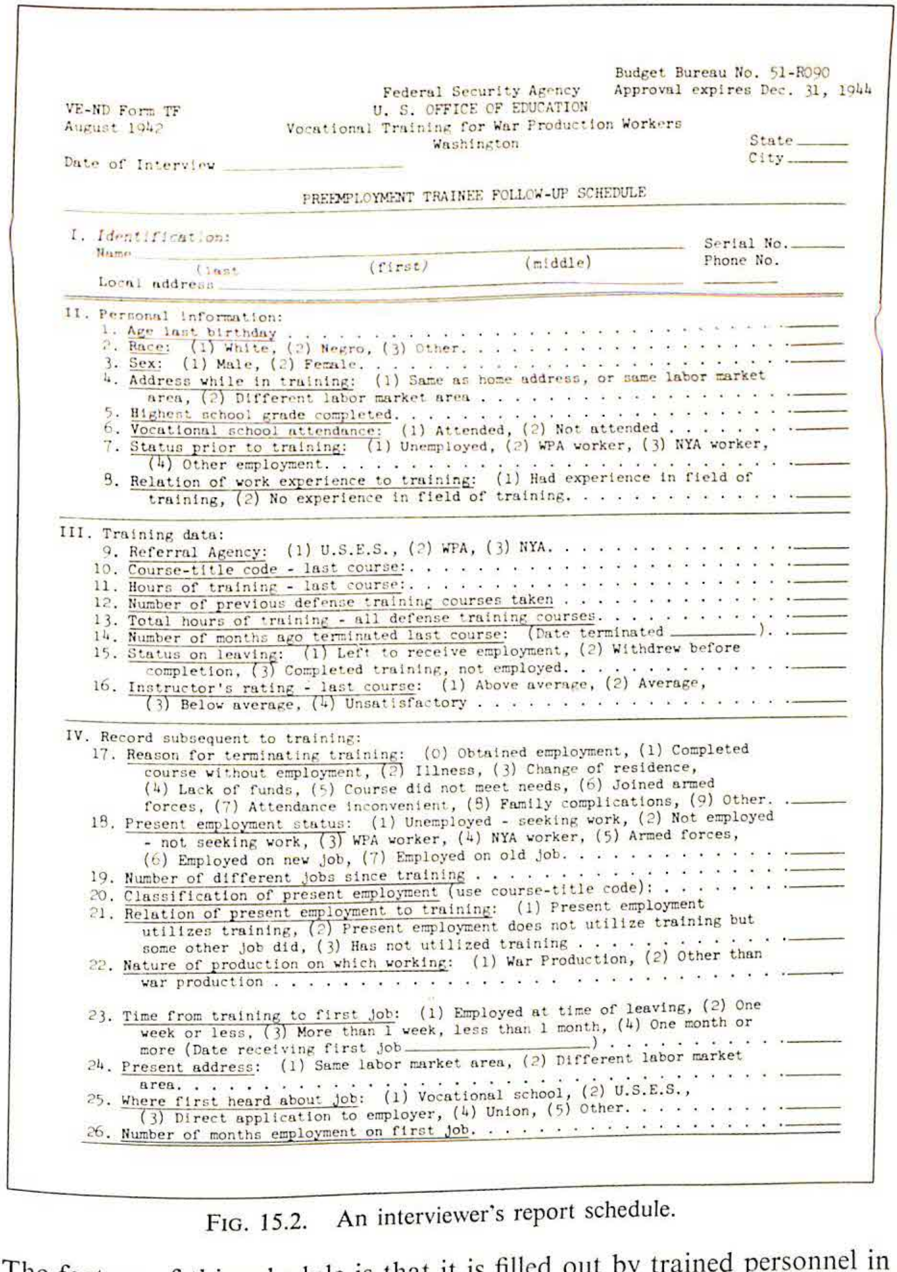
INFORMATION SUPPLIED BY:

Name .... *Title* ....

FIG. 15.1. Facsimile of a form satisfying criteria.

In Fig. 15.1 is reproduced the front of a return postcard used by the Office of Education in a sample survey of higher-education enrollments. The criteria listed above were applied in the development of the survey and of the form itself.

Figure 15.2 is a facsimile of an interviewer's report sheet used in a follow-up study of vocational war-production trainees during World War II. Several pages of uniform definitions are not reproduced here.



Budget Bureau No. 51-R090
Approval expires Dec. 31, 1944

VE-ND Form TF
August 1942

Federal Security Agency
U. S. OFFICE OF EDUCATION
Vocational Training for War Production Workers
Washington

State ______
City ______

Date of Interview ______________

PREEMPLOYMENT TRAINEE FOLLOW-UP SCHEDULE

I. *Identification:*

Name ________________________________ Serial No. ______
(last (first) (middle) Phone No.
Local address ________________________ ______

II. Personal information:
1. Age last birthday . . . . ______
2. Race: (1) White, (2) Negro, (3) Other. . . . . ______
3. Sex: (1) Male, (2) Female. . . . . ______
4. Address while in training: (1) Same as home address, or same labor market area, (2) Different labor market area . . . . ______
5. Highest school grade completed. . . . . ______
6. Vocational school attendance: (1) Attended, (2) Not attended . . . . ______
7. Status prior to training: (1) Unemployed, (2) WPA worker, (3) NYA worker, (4) Other employment. . . . . ______
8. Relation of work experience to training: (1) Had experience in field of training, (2) No experience in field of training. . . . . ______

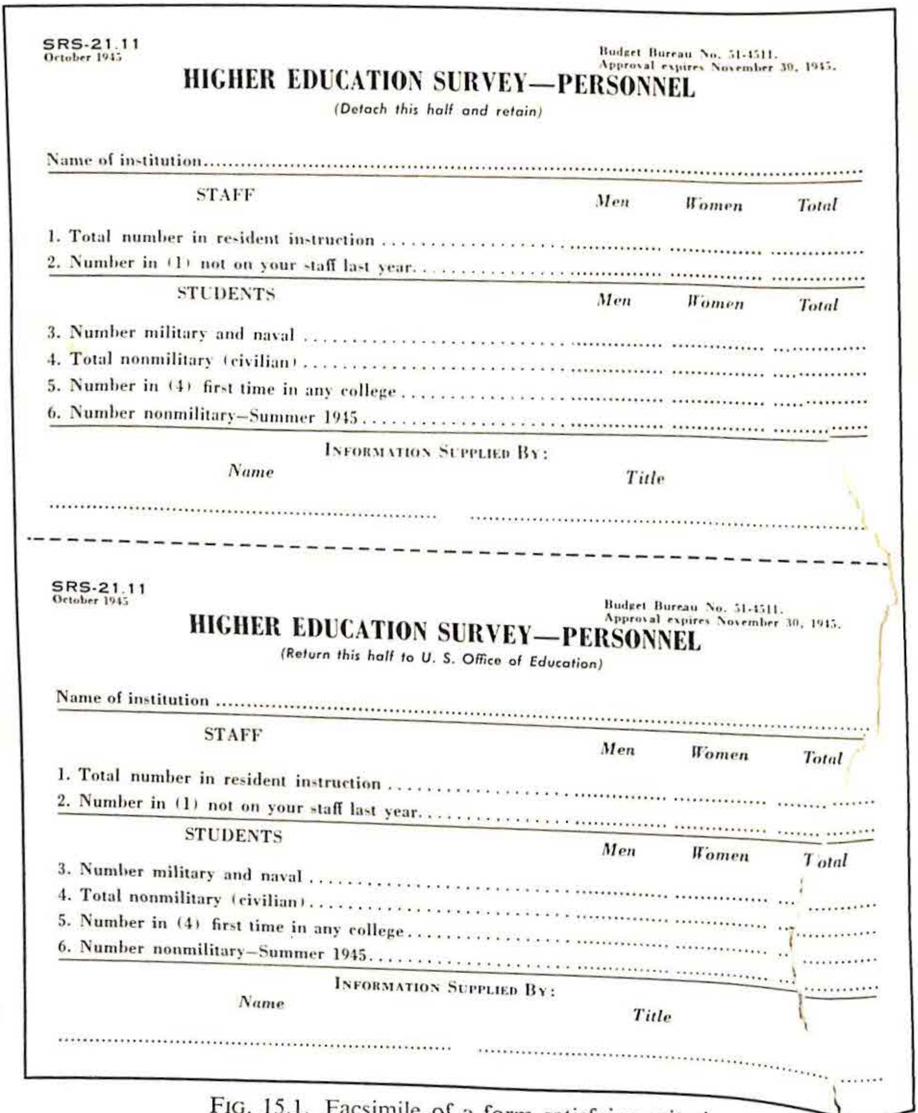
III. Training data:
9. Referral Agency: (1) U.S.E.S., (2) WPA, (3) NYA. . . . . ______
10. Course-title code - last course:. . . . . ______
11. Hours of training - last course:. . . . . ______
12. Number of previous defense training courses taken . . . . ______
13. Total hours of training - all defense training courses. . . . . ______
14. Number of months ago terminated last course: (Date terminated ______). . ______
15. Status on leaving: (1) Left to receive employment, (2) Withdrew before completion, (3) Completed training, not employed. . . . . ______
16. Instructor's rating - last course: (1) Above average, (2) Average, (3) Below average, (4) Unsatisfactory . . . . ______

IV. Record subsequent to training:
17. Reason for terminating training: (0) Obtained employment, (1) Completed course without employment, (2) Illness, (3) Change of residence, (4) Lack of funds, (5) Course did not meet needs, (6) Joined armed forces, (7) Attendance inconvenient, (8) Family complications, (9) Other. ______
18. Present employment status: (1) Unemployed - seeking work, (2) Not employed - not seeking work, (3) WPA worker, (4) NYA worker, (5) Armed forces, (6) Employed on new job, (7) Employed on old job. . . . . ______
19. Number of different jobs since training . . . . ______
20. Classification of present employment (use course-title code): . . . . ______
21. Relation of present employment to training: (1) Present employment utilizes training, (2) Present employment does not utilize training but some other job did, (3) Has not utilized training . . . . ______
22. Nature of production on which working: (1) War Production, (2) Other than war production . . . . ______
23. Time from training to first job: (1) Employed at time of leaving, (2) One week or less, (3) More than 1 week, less than 1 month, (4) One month or more (Date receiving first job ______) . . . . ______
24. Present address: (1) Same labor market area, (2) Different labor market area. . . . . ______
25. Where first heard about job: (1) Vocational school, (2) U.S.E.S., (3) Direct application to employer, (4) Union, (5) Other. . . . . ______
26. Number of months employment on first job. . . . . ______

FIG. 15.2. An interviewer's report schedule.

The feature of this schedule is that it is filled out by trained personnel in code form to minimize the work of mechanical tabulation which was part of the operating plan of the survey.[1]

[1] From *Preemployment Trainees in War Production*, U.S. Office of Education, Vocational Division Bulletin No. 224, Defense Training Series No. 2.



SRS-21.11
October 1945

Budget Bureau No. 51-4511.
Approval expires November 30, 1945.

HIGHER EDUCATION SURVEY—PERSONNEL

(Detach this half and retain)

Name of institution..

STAFF — *Men* — *Women* — *Total*

1. Total number in resident instruction ....
2. Number in (1) not on your staff last year.............. ....

STUDENTS — *Men* — *Women* — *Total*

3. Number military and naval ....
4. Total nonmilitary (civilian) ....
5. Number in (4) first time in any college ....
6. Number nonmilitary—Summer 1945 ....

INFORMATION SUPPLIED BY:

Name .... *Title* ....

SRS-21.11
October 1945

Budget Bureau No. 51-4511.
Approval expires November 30, 1945.

HIGHER EDUCATION SURVEY—PERSONNEL

(Return this half to U. S. Office of Education)

Name of institution ....

STAFF — *Men* — *Women* — *Total*

1. Total number in resident instruction ....
2. Number in (1) not on your staff last year.............. ....

STUDENTS — *Men* — *Women* — *Total*

3. Number military and naval ....
4. Total nonmilitary (civilian) ....
5. Number in (4) first time in any college ....
6. Number nonmilitary—Summer 1945 ....

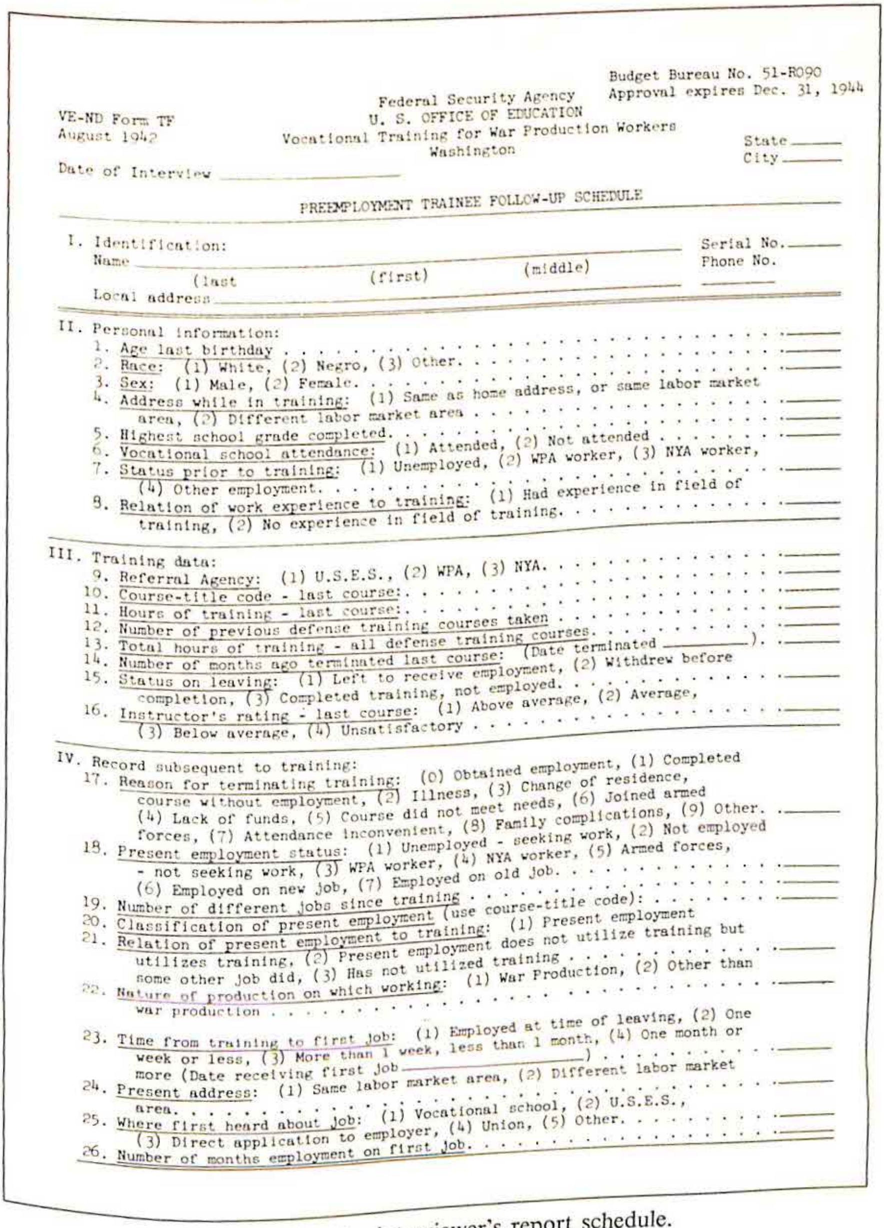
INFORMATION SUPPLIED BY:

Name .... *Title* ....

FIG. 15.1. Facsimile of a form satisfying criteria.

In Fig. 15.1 is reproduced the front of a return postcard used by the Office of Education in a sample survey of higher-education enrollments. The criteria listed above were applied in the development of the survey and of the form itself.

Figure 15.2 is a facsimile of an interviewer's report sheet used in a follow-up study of vocational war-production trainees during World War II. Several pages of uniform definitions are not reproduced here.



Budget Bureau No. 51-R090
Approval expires Dec. 31, 1944

VE-ND Form TF
August 1942

Federal Security Agency
U. S. OFFICE OF EDUCATION
Vocational Training for War Production Workers
Washington

State ______
City ______

Date of Interview ______

PREEMPLOYMENT TRAINEE FOLLOW-UP SCHEDULE

I. Identification:
Name ______ (last (first) (middle)
Local address ______
Serial No. ______
Phone No. ______

II. Personal information: . . . . ______
1. Age last birthday . . . . ______
2. Race: (1) White, (2) Negro, (3) Other. . . . . ______
3. Sex: (1) Male, (2) Female. . . . .
4. Address while in training: (1) Same as home address, or same labor market . . . ______
area, (2) Different labor market area . . . . ______
5. Highest school grade completed. . . . . ______
6. Vocational school attendance: (1) Attended, (2) Not attended . . . . ______
7. Status prior to training: (1) Unemployed, (2) WPA worker, (3) NYA worker,
(4) Other employment. ______
8. Relation of work experience to training: (1) Had experience in field of
training, (2) No experience in field of training. . . . . ______

III. Training data: . . . . ______
9. Referral Agency: (1) U.S.E.S., (2) WPA, (3) NYA. . . . . ______
10. Course-title code - last course:. . . . . ______
11. Hours of training - last course:. . . . . ______
12. Number of previous defense training courses taken . . . . ______
13. Total hours of training - all defense training courses. . . . . ______
14. Number of months ago terminated last course: (Date terminated ______). . ______
15. Status on leaving: (1) Left to receive employment, (2) Withdrew before
completion, (3) Completed training, not employed. . . . . ______
16. Instructor's rating - last course: (1) Above average, (2) Average,
(3) Below average, (4) Unsatisfactory . . . . ______

IV. Record subsequent to training:
17. Reason for terminating training: (0) Obtained employment, (1) Completed
course without employment, (2) Illness, (3) Change of residence,
(4) Lack of funds, (5) Course did not meet needs, (6) Joined armed
forces, (7) Attendance inconvenient, (8) Family complications, (9) Other. . ______
18. Present employment status: (1) Unemployed - seeking work, (2) Not employed
- not seeking work, (3) WPA worker, (4) NYA worker, (5) Armed forces,
(6) Employed on new job, (7) Employed on old job. . . . . ______
19. Number of different jobs since training . . . . ______
20. Classification of present employment (use course-title code): . . . . . . ______
21. Relation of present employment to training: (1) Present employment
utilizes training, (2) Present employment does not utilize training but
some other job did, (3) Has not utilized training . . . . ______
22. Nature of production on which working: (1) War Production, (2) Other than
war production . . . . ______
23. Time from training to first job: (1) Employed at time of leaving, (2) One
week or less, (3) More than 1 week, less than 1 month, (4) One month or
more (Date receiving first job ______) . . . . ______
24. Present address: (1) Same labor market area, (2) Different labor market
area. . . . . ______
25. Where first heard about job: (1) Vocational school, (2) U.S.E.S.,
(3) Direct application to employer, (4) Union, (5) Other. . . . . ______
26. Number of months employment on first job. . . . . ______

FIG. 15.2. An interviewer's report schedule.

The feature of this schedule is that it is filled out by trained personnel in code form to minimize the work of mechanical tabulation which was part of the operating plan of the survey.[1]

[1] From *Preemployment Trainees in War Production*, U.S. Office of Education, Vocational Division Bulletin No. 224, Defense Training Series No. 2.

15.2 ARRANGING DATA IN TABULAR FORM

We now come to a subject which does not require a high level of achievement in statistical theory. In fact, the problem of arranging data effectively in tables is as much an art as it is a science.

Through the statistical table the statistical worker seeks to communicate the results of his work to others. Unless he consciously exerts an effort to this task of communication, he is almost certain to fail at it. A helpful point of view in preparing tables is to put ourselves in the position of the individual on the receiving end of the communication channel, and looking at tabular presentation as a device for "telling a story."

It goes without saying that inaccuracy is to be avoided. A figure on total enrollment published in a state report is taken as authoritative, and the individual publishing the report has presumably taken all of the precautions in the handling of approximate figures and has taken precautionary steps in the checking and verifying of tabulations so as to eliminate or at least minimize clerical errors. Where there are doubts as to accuracy of the material in the table itself, this information should be indicated by footnotes in the table or as a part of the table, or at least in the text accompanying the table.

There is considerable latitude of choice in style of tables. Various publishers specify their own requirements on this subject. Among the agencies which publish tables in large volume should be mentioned the Federal government. The *Government Style Manual*[1] published by the Government Printing Office contains in some detail explicit rules regarding the make-up of tables. The U.S. Census, which publishes many volumes of statistical tables, has issued an extensive volume on tabular presentation.[2] One set of suggestions on tabular presentation, less detailed than the foregoing references, was prepared by Walker and Durost.[3] largely with educational statistics in mind. The following outline of points on tabular presentation is extracted in large part from Walker and Durost.

1. *The title.*

   a. A title should be completely unambiguous.

   b. All the information necessary for reading a table should appear with the table, so that the table can be lifted completely out of the context and still be understood.

[1] U.S. Government Printing Office, *Style Manual*, Washington, D.C., U.S. Government Printing Office, 1945, revised edition.

[2] Bruce L. Jenkinson, *Bureau of the Census Manual of Tabular Presentation*, Washington, D.C., U.S. Government Printing Office, 1949.

[3] Helen M. Walker and Walter M. Durost, *Statistical Tables, Their Structures and Use*, New York, Teachers College, Columbia University, 1936.

c. It is highly desirable to have a title as brief as is consistent with the necessity of imparting full information.

d. The usual rules of punctuation, capitals and lower case, should apply. Whatever form is selected should be uniformly adopted within an organization to simplify steps in preparation of tables.

e. It should always be remembered that a table is a list, and the title should name the item or items which the table lists.

f. Whenever a table is prepared individually without the support or amplification of any accompanying text, there should be included the appropriate identification of the institution, bureau, agency, or individual preparing the table, and the date.

2. *Stub and column headings.*

a. In general, every column should have a heading, and this heading should refer to the subheadings or to the tabulation of data directly under it.

b. Every row of the table should have a heading (in the "stub" column, usually at its extreme left).

c. The heading of the stub column should refer only to the items of the stub, and should not be used to describe the row in which the column headings stand.

d. The stub column usually appears at the extreme left of a table. Items in the stub identify data appearing in the columns of the table, row by row.

e. All subheads should derive logically from the box heading under which they stand, analysis proceeding from the more inclusive boxhead down to the separate column heads or line heads assigned to the individual arrays.

f. All the subheads under one box heading should be parallel in structure.

g. In a large table, rows and columns may be numbered to facilitate reference.

h. When headings must be long phrases, they can be accommodated better in the stub than in the columns.

i. When there is a quantitative classification, this is usually placed in the stub, and the numerical descriptions of class intervals serve as row headings.

j. Column headings are usually in the singular.

k. Categories should be mutually exclusive insofar as possible.

l. Categories should be all-inclusive insofar as possible.

3. *Table content.*

a. The term "total" should be reserved for those summaries which are obtained by the *addition* of other entries.

b. If a table extends over several pages, any footnote applying to the table as a whole is written on the last page. On any preceding page where there is an entry referring to the particular footnote, the same footnote symbol is inserted in the table and reference made at the bottom of page to footnote appearing on last page.

c. As in other aspects of table making, the main consideration determining the most appropriate order of entries is the function which the table is intended to serve.

d. In a frequency distribution it is usually best to arrange the classes used for tabulation in descending order, that is, from the highest class to the lowest.

e. A mixture of symbols, abbreviations, and words in the same table is not recommended.

f. Decimal points should be aligned in the columns.

g. The dollar sign is placed only before the first item in a column and before the total.

h. When *percent* is used as a column heading, the % sign is not placed after each entry in the column.

i. Numbers are rounded in accordance with the usual rule for such work.

j. A zero should be used to indicate a meaningful entry of zero size, *never to indicate lack of data or absence of frequency.*

k. Absence of data due to inability to secure returns, to loss of records, or the like, should be clearly indicated. Sometimes this is done with a dash in the appropriate cell of the table and an explanation in a footnote. Sometimes a separate category such as "no returns" is provided. A zero or an unexplained blank in such cases would be misleading.

l. To show that a percentage is not truly zero, yet too small to be recorded, a footnote symbol is inserted in the table, and the explanation is supplied in the footnote.

m. If data in a table are drawn from publications of another individual or agency, that fact should be made clear by complete reference to the source, including the place and date of publication, as well as the name of the author to whom they are to be credited.

n. Avoid splitting tables to extend over two pages. In preparing manuscripts for publication it is best to allow a separate page for each table.

4. *Ruling.*

a. The minimum ruling on a table is three horizontal lines; one above the body of the table, another below the body of the table, and a third separating the column headings from the actual data in the table.

b. When a table contains only two or three columns which can be separated by wide spacing, the use of vertical rules serves no purpose which cannot be served as well by the white space itself.

c. When a table crowds the page, so that it is not possible to provide enough space between the figures to let the eye select with certainty the figure for which it is looking, rules must be used.

d. In a complex table, it is often effective to use rules to mark off the major divisions only, using only white space to separate the items within the major divisions.

e. Inside a table, horizontal rules should be used only when there is a break in the structure to be emphasized, or a major subclassification be be set off. To use horizontal rules with every row of items is confusing and altogether undesirable.

f. Whatever plan of ruling is adopted, it should facilitate the kind of reading most appropriate for the table.

g. It is usually best to use horizontal ruling in tables constituting parts of forms or schedules to be filled in by other persons.

Table 15.1 is an example of one table as submitted by an inexperienced student. Table 15.2 is the same table, after revision. A good way to review the basic ideas about tables would be to list the things wrong with the first table.

TABLE 15.1

RESULTS OF THE OPINION POLLS

	TEACHERS							PUBLIC					
		P		C		NO.	%	P		C		NO.	%
1	8	368	95.3%	9	2.3%	386	100%	97	71.9%	35	25.9%	135	100%
2	9	376	97.4%	4	1%	386	100%	84	62.2%	45	33.3%	135	100%
3	10	337	87.3%	27	7%	386	100%	72	53.3%	58	43%	135	100%
4	11	384	99.5%			386	100%	96	71.1%	35	25.9%	135	100%
5	12	346	89.6%	12	3.1%	386	100%	64	47.4%	68	50.4%	135	100%
6	13	382	98.9%	2	.5%	386	100%	119	88.1%	15	11.1%	135	100%
7	14	377	97.6%	6	1.6%	366	100%	101	74.8%	31	23%	135	100%
8	15	377	97.6%	4	1%	386	100%	111	82.2%	20	14.8%	135	100%
9	16	372	96.3%	6	1.6%	386	100%	90	66.7%	39	28.9%	135	100%
10	17	340	88.1%	19	4.9%	386	100%	57	42.2%	72	53.3%	135	100%
11	18	337	87.3%	17	4.4%	386	100%	74	54.8%	52	38.5%	135	100%
12	19	349	90.4%	19	4.9%	386	100%	63	46.7%	70	51.9%	135	100%
13	20	384	99.5%	1	.3%	386	100%	113	83.7%	17	12.6%	135	100%
14	21	295	76.4%	59	15.3%	386	100%	20	14.8%	111	82.2%	135	100%
15	22	273	96.6%	2	.5%	386	100%	133	98.5%	1	.7%	135	100%
16	23	380	98.4%	3	.8%	386	100%	98	72.6%	31	23%	135	100%

TABLE 15.2

COMPARISON OF TEACHERS AND SAMPLE OF PUBLIC IN COMMUNITY X IN PROGRESSIVENESS OF RESPONSE TO SIXTEEN QUESTIONS OF EDUCATIONAL POLICY

Question	Progressive Responses			
	386 Teachers		135 Citizens	
	Number	Percent	Number	Percent
1. Teacher's role in classroom group	368	95.3	97	71.9
2. Classroom control and regulation	376	97.4	84	62.2
3. Deciding what to study	337	87.3	72	53.3
4. The teacher and pupil initiative	384	99.5	96	71.1
5. Group activity in school	346	89.6	64	47.4
6. Student participation in class discussion	382	98.9	119	88.1
7. Uniformity of class assignments	377	97.6	101	74.8
8. Work outside the classroom	377	97.6	111	82.2
9. The teacher and instructional materials	372	96.3	90	66.7
10. Racial, religious, and class differences	340	88.1	57	42.2
11. Life problems versus subject-matter fields	337	87.3	74	54.8
12. Marking and measuring pupil progress	349	90.4	63	46.7
13. All-round development of pupils	384	99.5	113	83.7
14. Elementary school promotion policy	295	76.4	20	14.8
15. Standards in the high school	273	96.6	133	98.5
16. Discipline in the school	380	98.4	98	72.6

REFERENCES

1. Dixon, Wilfrid J., and Frank J. Massey, *Introduction to Statistical Analysis*, New York, McGraw-Hill Book Co., 1951, Chapter 15.
2. Jenkinson, Bruce L., *Bureau of the Census Manual of Tabular Presentation*, Washington, D.C., U.S. Government Printing Office, 1949.
3. Neiswanger, William A., *Elementary Statistical Methods*, New York, The Macmillan Co., 1943, Chapters 1, 2, and 3.
4. Walker, Helen M., *Elementary Statistical Methods*, New York, Henry Holt and Co., 1943, Chapter 4.
5. Walker, Helen M., and Walter M. Durost, *Statistical Tables, Their Structures and Use*, New York, Teachers College, Columbia University, 1936.

APPENDIX A

Scores on Two Tests from a Midwest High School

The following scores are from the files of a Statewide Testing Program. They appear in 168 pairs. The first of each pair is a pupil's score on the California Test of Mental Maturity, X_1. The second is a test on Writing Skills, X_2, for the same pupil. The scores are respectively 96 and 56 for the first pupil.

X_1	X_2	X_1	X_2	X_1	X_2	X_1	X_2
96	56	48	45	64	48	65	27
67	43	79	47	84	52	82	54
81	54	85	56	79	36	71	42
49	31	76	47	83	59	61	46
93	54	69	54	61	48	77	59
82	47	88	53	79	48	98	56
58	50	84	49	70	56	88	57
60	40	91	54	69	38	53	46
71	42	65	40	79	50	68	47
51	34	59	47	51	39	50	30
61	35	67	57	96	57	78	40
74	34	89	52	58	28	78	51
70	45	70	42	76	51	52	53
93	59	81	59	98	60	70	32
51	41	65	49	72	44	60	29
74	44	82	56	49	18	73	38
76	48	61	36	72	50	54	25
78	47	80	50	73	42	95	60
73	24	62	28	72	43	65	48
51	31	67	48	62	45	59	37
61	47	58	38	82	51	58	46
64	27	78	58	77	56	78	53
60	55	88	54	71	55	34	24

X_1	X_2	X_1	X_2	X_1	X_2	X_1	X_2
61	29	56	30	62	37	41	25
74	51	77	45	70	45	73	30
61	32	61	47	76	32	83	33
66	36	62	39	79	49	65	56
90	58	67	41	95	57	65	46
74	43	77	49	72	53	60	50
52	37	85	55	46	45	36	29
40	32	76	41	44	40	50	36
79	47	81	58	64	46	85	45
71	25	68	56	78	57	73	51
87	44	80	60	74	24	59	36
72	48	58	21	59	48	85	56
77	49	77	48	76	48	51	47
74	51	73	49	73	15	64	47
72	29	64	29	79	51	92	56
95	58	86	48	78	48	71	53
75	40	72	44	77	51	54	40
53	40	64	46	84	49	64	47
41	48	55	32	70	49	63	31

APPENDIX B

A Table of Random Numbers

Reproduced through the courtesy of the Rand Corporation and the *Journal of the American Statistical Association* 49: 683-84, September 1954, from *A Million Random Digits*, Glencoe, Illinois, The Free Press, 1955.

58932	89326	33491	04617	88092
73073	52171	89301	74066	82717
42665	80748	77622	15779	37361
59985	59807	60562	85747	94028
50943	40422	63035	60344	06883
22224	02627	91576	16781	89184
24473	42096	76920	88864	54164
38582	21871	14672	93362	67981
46094	43845	91838	79574	08003
91061	31674	73729	99315	16699
00397	56753	53158	71872	68153
14328	44708	72952	27048	67887
88534	87112	68614	83073	88794
97347	87316	73087	77135	71883
01366	72976	01868	51667	63279
37106	20523	21584	93712	83654
06476	70603	97122	44978	78028
81717	48410	94516	15427	85323
51583	69788	41758	55004	30992
50120	33884	83655	88345	69602
89761	23053	77480	28683	68324
08943	66660	11057	98849	29499
71685	97247	79368	43710	80365
17402	66300	94385	01717	96191
52606	39860	92127	42588	93307
66035	07223	76264	29148	68652
21565	30786	45403	33782	93424
88735	75275	03080	77653	55430
50404	80166	28017	52611	60012
80834	11317	93109	91857	47904
26872	72927	79021	51571	68825
16530	96086	17329	87959	23727
84644	00448	86828	50552	84832
88620	72894	94716	84622	49771
22209	78590	68615	58113	23727
04795	53971	14592	39634	03855
54291	56045	61635	32186	86651
30654	48543	18339	65024	33386
11123	08732	49393	12911	75803
56577	51257	83291	12329	17827
58987	02026	42969	59144	84349
16851	99197	70476	77113	46320
02104	49435	77706	18924	24957
54440	07893	31618	35707	65130
87681	42543	69847	81848	32034
24337	61634	52574	83649	28725
62557	25292	72781	17186	10393
02913	03885	58822	82941	43806
68706	87619	13846	56197	27151
05930	33213	78416	00194	91369

70119	31347	12659	11574	70052
98390	30240	28330	41145	16918
08172	23823	48433	57222	34435
21238	19051	50768	40807	88681
79342	44640	93942	97371	16842
93039	79367	00812	41365	04515
62865	09576	97207	33739	78345
00800	72496	24767	61768	07228
64340	02224	48336	14891	72188
92168	52692	31224	12185	43065
20494	18813	16242	40257	66402
87693	30242	70545	69128	51528
05567	05561	82071	07234	67690
85166	37189	75671	33879	27411
26704	47922	56650	40236	66207
01047	81634	77395	62310	41501
58183	21952	84098	28913	55736
64667	57092	21315	04731	71877
27149	13843	09817	09407	88276
66232	80293	74502	36925	60184
40500	21406	00571	87320	81683
35892	49668	83991	72088	30210
54819	26094	51409	21485	94764
64224	47909	09994	23750	17351
36913	58173	45709	83679	82617
64254	64745	10614	86371	43244
82018	25536	74031	31807	70133
28833	44043	96215	21270	59427
96879	27659	95463	53847	40921
95938	76014	99818	16606	19713
97154	71237	06073	57343	51428
78790	17026	59008	28543	11576
25034	59325	08844	95774	49323
70116	44091	88505	15575	44927
66904	23000	73259	68626	98962
91171	28299	62619	81550	46798
74547	13260	79262	55831	83784
30448	14154	75795	39465	82353
06584	29867	45898	66415	89349
68548	86576	14344	75889	04514
49319	50206	22024	56124	50749
81034	86779	34622	70859	33045
68905	44234	18244	31602	38388
88530	72096	44459	31449	93182
37227	11302	04667	32526	64713
83220	50529	20619	11606	10297
66703	30017	35347	35038	16648
69556	76728	60535	59961	76979
99040	96390	65989	38375	30332
85185	72849	58611	31220	66108

APPENDIX C

Ordinates and Areas of the Normal Curve*

z	$f(X)$	$\int_0^z$	z	$f(X)$	$\int_0^z$	z	$f(X)$	$\int_0^z$
.00	.39894	.00000	.30	.38139	.11791	.60	.33322	.22575
.01	.39892	.00399	.31	.38023	.12172	.61	.33121	.22907
.02	.39886	.00798	.32	.37903	.12552	.62	.32918	.23237
.03	.39876	.01197	.33	.37780	.12930	.63	.32713	.23565
.04	39862	.01595	.34	.37654	.13307	.64	.32506	.23891
.05	.39844	.01994	.35	.37524	.13683	.65	.32297	.24215
.06	.39822	.02392	.36	.37391	.14058	.66	.32086	.24537
.07	.39797	.02790	.37	.37255	.14431	.67	.31874	.24857
.08	.39767	.03188	.38	.37115	.14803	.68	.31659	.25175
.09	.39733	.03586	.39	.36973	.15173	.69	.31443	.25490
.10	.39695	.03983	.40	.36827	.15542	.70	.31225	.25804
.11	.39654	.04380	.41	.36678	.15910	.71	.31006	.26115
.12	.39608	.04776	.42	.36526	.16276	.72	.30785	.26424
.13	.39559	.05172	.43	.36371	.16640	.73	.30563	.26730
.14	.39505	.05567	.44	.36213	.17003	.74	.30339	.27035
.15	.39448	.05962	.45	.36053	.17364	.75	.30114	.27337
.16	.39387	.06356	.46	.35889	.17724	.76	.29887	.27637
.17	.39322	.06749	.47	.35723	.18082	.77	.29659	.27935
.18	.39253	.07142	.48	.35553	.18439	.78	.29431	.28230
.19	.39181	.07535	.49	.35381	.18793	.79	.29200	.28524
.20	.39104	.07926	.50	.35207	.19146	.80	.28969	.28814
.21	.39024	.08317	.51	.35029	.19497	.81	.28737	.29103
.22	.38940	.08706	.52	.34849	.19847	.82	.28504	.29389
.23	.38853	.09095	.53	.34667	.20194	.83	.28269	.29673
.24	.38762	.09483	.54	.34482	.20540	.84	.28034	.29955
.25	.38667	.09871	.55	.34294	.20884	.85	.27798	.30234
.26	.38568	.10257	.56	.34105	.21226	.86	.27562	.30511
.27	.38466	.10642	.57	.33912	.21566	.87	.27324	.30785
.28	.38361	.11026	.58	.33718	.21904	.88	.27086	.31057
.29	.38251	.11409	.59	.33521	.22240	.89	.26848	.31327

[*] This table reproduced from *Mathematics of Statistics*, by John F. Kenney, with permission of the publishers D. Van Nostrand Company, Inc.

Ordinates and Areas of the Normal Curve (*continued*)

z	$f(X)$	$\int_0^z$	z	$f(X)$	$\int_0^z$	z	$f(X)$	$\int_0^z$
.90	.26609	.31594	1.25	.18265	.39435	1.60	.11092	.44520
.91	.26369	.31859	1.26	.18037	.39617	1.61	.10915	.44630
.92	.26129	.32121	1.27	.17810	.39796	1.62	.10741	.44738
.93	.25888	.32381	1.28	.17585	.39973	1.63	.10567	.44845
.94	.25647	.32639	1.29	.17360	.40147	1.64	.10396	.44950
.95	.25406	.32894	1.30	.17137	.40320	1.65	.10226	.45053
.96	.25164	.33147	1.31	.16915	.40490	1.66	.10059	.45154
.97	.24923	.33398	1.32	.16694	.40658	1.67	.09893	.45254
.98	.24681	.33646	1.33	.16474	.40824	1.68	.09728	.45352
.99	.24439	.33891	1.34	.16256	.40988	1.69	.09566	.45449
1.00	.24197	.34134	1.35	.16038	.41149	1.70	.09405	.45543
1.01	.23955	.34375	1.36	.15822	.41309	1.71	.09246	.45637
1.02	.23713	.34614	1.37	.15608	.41466	1.72	.09089	.45728
1.03	.23471	.34850	1.38	.15395	.41621	1.73	.08933	.45818
1.04	.23230	.35083	1.39	.15183	.41774	1.74	.08780	.45907
1.05	.22988	.35314	1.40	.14973	.41924	1.75	.08628	.45994
1.06	.22747	.35543	1.41	.14764	.43073	1.76	.08478	.46080
1.07	.22506	.35769	1.42	.14556	.42220	1.77	.08329	.46164
1.08	.22265	.35993	1.43	.14350	.42364	1.78	.08183	.46246
1.09	.22025	.36214	1.44	.14146	.42507	1.79	.08038	.46327
1.10	.21785	.36433	1.45	.13943	.42647	1.80	.07895	.46407
1.11	.21546	.36650	1.46	.13742	.42786	1.81	.07754	.46485
1.12	.21307	.36864	1.47	.13542	.42922	1.82	.07614	.46562
1.13	.21069	.37076	1.48	.13344	.43056	1.83	.07477	.46638
1.14	.20831	.37286	1.49	.13147	.43189	1.84	.07341	.46712
1.15	.20594	.37493	1.50	.12952	.43319	1.85	.07206	.46784
1.16	.20357	.37698	1.51	.12758	.43448	1.86	.07074	.46856
1.17	.20121	.37900	1.52	.12566	.43574	1.87	.06943	.46926
1.18	.19886	.38100	1.53	.12376	.43699	1.88	.06814	.46995
1.19	.19652	.38298	1.54	.12188	.43822	1.89	.06687	.47062
1.20	.19419	.38493	1.55	.12001	.43943	1.90	.06562	.47128
1.21	.19186	.38686	1.56	.11816	.44062	1.91	.06439	.47193
1.22	.18954	.38877	1.57	.11632	.44179	1.92	.06316	.47257
1.23	.18724	.39065	1.58	.11450	.44295	1.93	.06195	.47320
1.24	.18494	.39251	1.59	.11270	.44408	1.94	.06077	.47381

Ordinates and Areas of the Normal Curve (*continued*)

z	$f(X)$	$\int_0^z$	z	$f(X)$	$\int_0^z$	z	$f(X)$	$\int_0^z$
1.95	.05959	.47441	2.30	.02833	.48928	2.65	.01191	.49598
1.96	.05844	.47500	2.31	.02768	.48956	2.66	.01160	.49609
1.97	.05730	.47558	2.32	.02705	.48983	2.67	.01130	.49621
1.98	.05618	.47615	2.33	.02643	.49010	2.68	.01100	.49632
1.99	.05508	.47670	2.34	.02582	.49036	2.69	.01071	.49643
2.00	.05399	.47725	2.35	.02522	.49061	2.70	.01042	.49653
2.01	.02592	.47778	2.36	.02463	.49086	2.71	.01014	.49664
2.02	.05186	.47831	2.37	.02406	.49111	2.72	.00987	.49674
2.03	.05082	.47882	2.38	.02349	.49134	2.73	.00961	.49683
2.04	.04980	.47932	2.39	.02294	.49158	2.74	.00935	.49693
2.05	.04879	.47982	2.40	.02239	.49180	2.75	.00909	.49702
2.06	.04780	.48030	2.41	.02186	.49202	2.76	.00885	.49711
2.07	.04682	.48077	2.42	.02134	.49224	2.77	.00861	.49720
2.08	.04586	.48124	2.43	.02083	.49245	2.78	.00837	.49728
2.09	.04491	.48169	2.44	.02033	.49266	2.79	.00814	.49736
2.10	.04398	.48214	2.45	.01984	.49286	2.80	.00792	.49744
2.11	.04307	.48257	2.46	.01936	.49305	2.81	.00770	.49752
2.12	.04217	.48300	2.47	.01889	.49324	2.82	.00748	.49760
2.13	.04128	.48341	2.48	.01842	.49343	2.83	.00727	.49767
2.14	.04041	.48382	2.49	.01797	.49361	2.84	.00707	.49774
2.15	.03955	.48422	2.50	.01753	.49379	2.85	.00687	.49781
2.16	.03871	.48461	2.51	.01709	.49396	2.86	.00668	.49788
2.17	.03788	.48500	2.52	.01667	.49413	2.87	.00649	.49795
2.18	.03706	.48537	2.53	.01625	.49430	2.88	.00631	.49801
2.19	.03626	.48574	2.54	.01585	.49446	2.89	.00613	.49807
2.20	.03547	.48610	2.55	.01545	.49461	2.90	.00595	.49813
2.21	.03470	.48645	2.56	.01506	.49477	2.91	.00578	.49819
2.22	.03394	.48679	2.57	.01468	.49492	2.92	.00562	.49825
2.23	.03319	.48713	2.58	.01431	.49506	2.93	.00545	.49831
2.24	.03246	.48745	2.59	.01394	.49520	2.94	.00530	.49836
2.25	.03174	.48778	2.60	.01358	.49534	2.95	.00514	.49841
2.26	.03103	.48809	2.61	.01323	.49547	2.96	.00499	.49846
2.27	.03034	.48840	2.62	.01289	.49560	2.97	.00485	.49851
2.28	.02965	.48870	2.63	.01256	.49573	2.98	.00471	.49856
2.29	.02898	.48899	2.64	.01223	.49585	2.99	.00457	.49861

Ordinates and Areas of the Normal Curve (*continued*)

z	$f(X)$	$\int_0^z$	z	$f(X)$	$\int_0^z$	z	$f(X)$	$\int_0^z$
3.00	.00443	.49865	3.35	.00146	.49960	3.70	.00042	.49989
3.01	.00430	.49869	3.36	.00141	.49961	3.71	.00041	.49990
3.02	.00417	.49874	3.37	.00136	.49962	3.72	.00039	.49990
3.03	.00405	.49878	3.38	.00132	.49964	3.73	.00038	.49990
3.04	.00393	.49882	3.39	.00127	.49965	3.74	.00037	.49991
3.05	.00381	.49886	3.40	.00123	.49966	3.75	.00035	.49991
3.06	.00370	.49889	3.41	.00119	.49968	3.76	.00034	.49992
3.07	.00358	.49893	3.42	.00115	.49969	3.77	.00033	.49992
3.08	.00348	.49897	3.43	.00111	.49970	3.78	.00031	.49992
3.09	.00337	.49900	3.44	.00107	.49971	3.79	.00030	.49992
3.10	.00327	.49903	3.45	.00104	.49972	3.80	.00029	.49993
3.11	.00317	.49906	3.46	.00100	.49973	3.81	.00028	.49993
3.12	.00307	.49910	3.47	.00097	.49974	3.82	.00027	.49993
3.13	.00298	.49913	3.48	.00094	.49975	3.83	.00026	.49994
3.14	.00288	.49916	3.49	.00090	.49976	3.84	.00025	.49994
3.15	.00279	.49918	3.50	.00087	.49977	3.85	.00024	.49994
3.16	.00271	.49921	3.51	.00084	.49978	3.86	.00023	.49994
3.17	.00262	.49924	3.52	.00081	.49978	3.87	.00022	.49995
3.18	.00254	.49926	3.53	.00079	.49979	3.88	.00021	.49995
3.19	.00246	3.6429	3.54	.00076	.49980	3.89	.00021	.49995
3.20	.00238	.49931	3.55	.00073	.49981	3.90	.00020	.49995
3.21	.00231	.49934	3.56	.00071	.49981	3.91	.00019	.49995
3.22	.00224	.49936	3.57	.00068	.49982	3.92	.00018	.49996
3.23	.00216	.49938	3.58	.00066	.49983	3.93	.00018	.49996
3.24	.00210	.49940	3.59	.00063	.49983	3.94	.00017	.49996
3.25	.00203	.49942	3.60	.00061	.49984	3.95	.00016	.49996
3.26	.00196	.49944	3.61	.00059	.49985	3.96	.00016	.49996
3.27	.00190	.49946	3.62	.00057	.49985	3.97	.00015	.49996
3.28	.00184	.49948	3.63	.00055	.49986	3.98	.00014	.49997
3.29	.00178	.49950	3.64	.00053	.49986	3.99	.00014	.49997
3.30	.00172	.49952	3.65	.00051	.49987			
3.31	.00167	.49953	3.66	.00049	.49987			
3.32	.00161	.49955	3.67	.00047	.49988			
3.33	.00156	.49957	3.68	.00046	.49988			
3.34	.00151	.49958	3.69	.00044	.49989			

APPENDIX D

Cumulative Normal Distribution*

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

	1.282	1.645	1.960	2.326	2.576	3.090	3.291	3.891	4.417
$F(x)$	.90	.95	.975	.99	.995	.999	.9995	.99995	.999995
$2[1 - F(x)]$	.20	.10	.05	.02	.01	.002	.001	.0001	.00001

* Reproduced by permission from *Introduction to the Theory of Statistics*, by A. M. Mood. Copyright 1950. McGraw-Hill Book Company, Inc.

APPENDIX E

Distribution of Student's t

	$\alpha = .05$		$\alpha = .01$	
	(One-Tailed Test)	(Two-tailed Test)	(One-Tailed Test)	(Two-Tailed Test)
d.f.	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$
1	6.31	12.71	31.82	63.66
2	2.92	4.30	6.96	9.92
3	2.35	3.18	4.54	5.84
4	2.13	2.78	3.75	4.60
5	2.01	2.57	3.36	4.03
6	1.94	2.45	3.14	3.71
7	1.89	2.36	3.00	3.50
8	1.86	2.31	2.90	3.36
9	1.83	2.26	2.82	3.25
10	1.81	2.23	2.76	3.17
11	1.80	2.20	2.72	3.11
12	1.78	2.18	2.68	3.05
13	1.77	2.16	2.65	3.01
14	1.76	2.14	2.62	2.98
15	1.75	2.13	2.60	2.95
16	1.75	2.12	2.58	2.92
17	1.74	2.11	2.57	2.90
18	1.73	2.10	2.55	2.88
19	1.73	2.09	2.54	2.86
20	1.72	2.09	2.53	2.85
21	1.72	2.08	2.52	2.83
22	1.72	2.07	2.51	2.82
23	1.71	2.07	2.50	2.81
24	1.71	2.06	2.49	2.80
25	1.71	2.06	2.48	2.79
26	1.71	2.06	2.48	2.78
27	1.70	2.05	2.47	2.77
28	1.70	2.05	2.47	2.76
29	1.70	2.05	2.46	2.76
30	1.70	2.04	2.46	2.75
40	1.68	2.02	2.42	2.70
60	1.67	2.00	2.39	2.66
120	1.66	1.98	2.36	2.62
∞	1.64	1.96	2.33	2.58

APPENDIX F

Table for Transforming r to z' and Vice Versa

r	z'	r	z'	r	z'	r	z'	r	z'
.000	.000	.200	.203	.400	.424	.600	.693	.800	1.099
.005	.005	.205	.208	.405	.430	.605	.701	.805	1.113
.010	.010	.210	.213	.410	.436	.610	.709	.810	1.127
.015	.015	.215	.218	.415	.442	.615	.717	.815	1.142
.020	.020	.220	.224	.420	.448	.620	.725	.820	1.157
.025	.025	.225	.229	.425	.454	.625	.733	.825	1.172
.030	.030	.230	.234	.430	.460	.630	.741	.830	1.188
.035	.035	.235	.239	.435	.466	.635	.750	.835	1.204
.040	.040	.240	.245	.440	.472	.640	.758	.840	1.221
.045	.045	.245	.250	.445	.478	.645	.767	.845	1.238
.050	.050	.250	.255	.450	.485	.650	.775	.850	1.256
.055	.055	.255	.261	.455	.491	.655	.784	.855	1.274
.060	.060	.260	.266	.460	.497	.660	.793	.860	1.293
.065	.065	.265	.271	.465	.504	.665	.802	.865	1.313
.070	.070	.270	.277	.470	.510	.670	.811	.870	1.333
.075	.075	.275	.282	.475	.517	.675	.820	.875	1.354
.080	.080	.280	.288	.480	.523	.680	.829	.880	1.376
.085	.085	.285	.293	.485	.530	.685	.838	.885	1.398
.090	.090	.290	.299	.490	.536	.690	.848	.890	1.422
.095	.095	.295	.304	.495	.543	.695	.858	.895	1.447
.100	.100	.300	.310	.500	.549	.700	.867	.900	1.472
.105	.105	.305	.315	.505	.556	.705	.877	.905	1.499
.110	.110	.310	.321	.510	.563	.710	.887	.910	1.528
.115	.116	.315	.326	.515	.570	.715	.897	.915	1.557
.120	.121	.320	.332	.520	.576	.720	.908	.920	1.589
.125	.126	.325	.337	.525	.583	.725	.918	.925	1.623
.130	.131	.330	.343	.530	.590	.730	.929	.930	1.658
.135	.136	.335	.348	.535	.597	.735	.940	.935	1.697
.140	.141	.340	.354	.540	.604	.740	.950	.940	1.738
.145	.146	.345	.360	.545	.611	.745	.962	.945	1.783
.150	.151	.350	.365	.550	.618	.750	.973	.950	1.832
.155	.156	.355	.371	.555	.626	.755	.984	.955	1.886
.160	.161	.360	.377	.560	.633	.760	.996	.960	1.946
.165	.167	.365	.383	.565	.640	.765	1.008	.965	2.014
.170	.172	.370	.388	.570	.648	.770	1.020	.970	2.092
.175	.177	.375	.394	.575	.655	.775	1.033	.975	2.185
.180	.182	.380	.400	.580	.662	.780	1.045	.980	2.298
.185	.187	.385	.406	.585	.670	.785	1.058	.985	2.443
.190	.192	.390	.412	.590	.678	.790	1.071	.990	2.647
.195	.198	.395	.418	.595	.685	.795	1.085	.995	2.994

APPENDIX G

Selected Data, Four Illinois High Schools

The accompanying table contains extracts from reports filed in the office of the Illinois statewide high-school testing program. Among other information available for all high school seniors was:

(1) Sex of student.

(2) Whether or not student planned to attend college.

(3) Scores on the California Test of Mental Maturity.

(4) A physical science reading test.

The table gives essential information summarizing these data for four samples of seniors from four separate high schools.

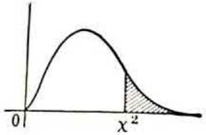
TABULAR SUMMARY OF BASIC DATA FROM FOUR HIGH SCHOOLS

School	Category of Student	n	ΣX[a]	ΣY[b]	ΣX^2	ΣY^2	ΣXY
H.S. A	Male—College	13	1,021	210	82,895	4,258	17,886
	Male—Noncollege	9	574	95	37,458	1,025	6,042
	Female—College	14	897	187	59,167	2,837	12,432
	Female—Noncollege	9	592	101	39,580	1,337	6,865
	School Total	45	3,084	593	219,100	9,457	43,225
H.S. B	Male—College	13	901	220	63,603	4,654	15,974
	Male—Noncollege	8	496	102	30,912	1,396	6,406
	Female—College	5	312	56	19,994	688	3,630
	Female—Noncollege	9	530	94	32,368	1,030	5,665
	School Total	35	2,239	472	146,877	7,768	31,675
H.S. C	Male—College	9	709	191	57,557	4,769	15,987
	Male—Noncollege	8	511	105	33,033	1,519	6,900
	Female—College	5	340	69	23,220	1,055	4,762
	Female—Noncollege	13	776	145	48,640	1,741	8,660
	School Total	35	2,336	510	162,450	9,084	36,309
H.S. D	Male—College	25	2,085	525	179,283	12,929	44,940
	Male—Noncollege	2	154	37	11,890	689	2,837
	Female—College	15	1,188	220	94,774	3,560	17,603
	Female—Noncollege	2	121	22	7,345	242	1,331
	School Total	44	3,548	804	293,292	17,420	66,711
All four	Male—College	60	4,716	1,146	383,338	26,610	94,787
	Male—Noncollege	27	1,735	339	113,293	4,629	22,185
	Female—College	39	2,737	532	197,155	8,140	38,427
	Female—Noncollege	33	2,019	362	127,933	4,350	22,521
	Grand Total	159	11,207	2,379	821,719	43,729	177,920

[a] CTMM. [b] Phys. Sci.

APPENDIX H

χ^2 Distribution*



Degrees of freedom	$P = 0.99$	0.98	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.02	0.01
1	0.000157	0.000628	0.00393	0.0158	0.0642	0.148	0.455	1.074	1.642	2.706	3.841	5.412	6.635
2	0.0201	0.0404	0.103	0.211	0.446	0.713	1.386	2.408	3.219	4.605	5.991	7.824	9.210
3	0.115	0.185	0.352	0.584	1.005	1.424	2.366	3.665	4.642	6.251	7.815	9.837	11.341
4	0.297	0.429	0.711	1.064	1.649	2.195	3.357	4.878	5.989	7.779	9.488	11.668	13.277
5	0.554	0.752	1.145	1.610	2.343	3.000	4.351	6.064	7.289	9.236	11.070	13.388	15.086
6	0.872	1.134	1.635	2.204	3.070	3.828	5.348	7.231	8.558	10.645	12.592	15.033	16.812
7	1.239	1.564	2.167	2.833	3.822	4.671	6.346	8.383	9.803	12.017	14.067	16.622	18.475
8	1.646	2.032	2.733	3.490	4.594	5.527	7.344	9.524	11.030	13.362	15.507	18.168	20.090
9	2.088	2.532	3.325	4.168	5.380	6.393	8.343	10.656	12.242	14.684	16.919	19.679	21.666
10	2.558	3.059	3.940	4.865	6.179	7.267	9.342	11.781	13.442	15.987	18.307	21.161	23.209
11	3.053	3.609	4.575	5.578	6.989	8.148	10.341	12.899	14.631	17.275	19.675	22.618	24.725
12	3.571	4.178	5.226	6.304	7.807	9.034	11.340	14.011	15.812	18.549	21.026	24.054	26.217
13	4.107	4.765	5.892	7.042	8.634	9.926	12.340	15.119	16.985	19.812	22.362	25.472	27.688
14	4.660	5.368	6.571	7.790	9.467	10.821	13.339	16.222	18.151	21.064	23.685	26.873	29.141
15	5.229	5.985	7.261	8.547	10.307	11.721	14.339	17.322	19.311	22.307	24.996	28.259	30.578
16	5.812	6.614	7.962	9.312	11.152	12.624	15.338	18.418	20.465	23.542	26.296	29.633	32.000
17	6.408	7.255	8.672	10.085	12.002	13.531	16.338	19.511	21.615	24.769	27.587	30.995	33.409
18	7.015	7.906	9.390	10.865	12.857	14.440	17.338	20.601	22.760	25.989	28.869	32.346	34.805
19	7.633	8.567	10.117	11.651	13.716	15.352	18.338	21.689	23.900	27.204	30.144	33.687	36.191
20	8.260	9.237	10.851	12.443	14.578	16.266	19.337	22.775	25.038	28.412	31.410	35.020	37.566
21	8.897	9.915	11.591	13.240	15.445	17.182	20.337	23.858	26.171	29.615	32.671	36.343	38.932
22	9.542	10.600	12.338	14.041	16.314	18.101	21.337	24.939	27.301	30.813	33.924	37.659	40.289
23	10.196	11.293	13.091	14.848	17.187	19.021	22.337	26.018	28.429	32.007	35.172	38.968	41.638
24	10.856	11.992	13.848	15.659	18.062	19.943	23.337	27.096	29.553	33.196	36.415	40.270	42.980
25	11.524	12.697	14.611	16.473	18.940	20.867	24.337	28.172	30.675	34.382	37.652	41.566	44.314
26	12.198	13.409	15.379	17.292	19.820	21.792	25.336	29.246	31.795	35.563	38.885	42.856	45.642
27	12.879	14.125	16.151	18.114	20.703	22.719	26.336	30.319	32.912	36.741	40.113	44.140	46.963
28	13.565	14.847	16.928	18.939	21.588	23.647	27.336	31.391	34.027	37.916	41.337	45.419	48.278
29	14.256	15.574	17.708	19.768	22.475	24.577	28.336	32.461	35.139	39.087	42.557	46.693	49.588
30	14.953	16.306	18.493	20.599	23.364	25.508	29.336	33.530	36.250	40.256	43.773	47.962	50.892

For degrees of freedom greater than 30, the expression $\sqrt{2\chi^2} - \sqrt{2n' - 1}$ may be used as a normal deviate with unit variance, where n' is the number of degrees of freedom.

*Reprinted from Table III of R. H. Fisher, *Statistical Methods for Research Workers;* published by Oliver & Boyd, Ltd., Edinburgh, by permission of the author and publisher.

APPENDIX I

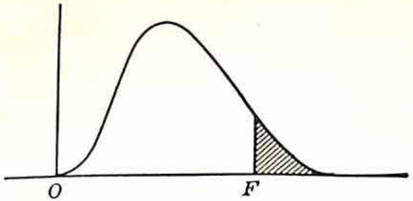
The following table consists of 95th percentile and 99th percentile values of F for various combinations of degrees of freedom. The 5 percent points (in roman type) and the 1 percent points (in bold-face type) are the points below which 95 percent and 99 percent of sample F ratios are expected respectively. These points are thus critical values for *one-sided* tests.

Testing the homogeneity of variance of two samples is a two-sided test. For such a test the larger variance is placed in the numerator of the F ratio and the tabular probability is doubled. The values of F shown in the table are thus, for such purposes, 10 percent and 5 percent critical values. To find 1 percentile and 5 percentile values of F, see pages 233–234.

Ordinarily in using this table, enter first across the top, selecting the proper column for the number of degrees of freedom in the *numerator*. Then proceed down this column until the row for the number of degrees of freedom in the denominator is reached. This locates the cell in which appears the two *one-sided critical values*, $F_{.95}$ and $F_{.99}$.

APPENDIX I

F Distribution*



5% (Roman Type) and 1% (Bold-Face Type) Points for the Distribution of *F*

Degrees of freedom for lesser mean square	Degrees of freedom for greater mean square																						
	1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	40	50	75	100	200	500	∞
1	161 **4052**	200 **4999**	216 **5403**	225 **5625**	230 **5764**	234 **5859**	237 **5928**	239 **5981**	241 **6022**	242 **6056**	243 **6082**	244 **6106**	245 **6142**	246 **6169**	248 **6208**	249 **6234**	250 **6258**	251 **6286**	252 **6302**	253 **6323**	253 **6334**	254 **6352**	254 **6361**	254 **6366**
2	18.51 **98.49**	19.00 **99.01**	19.16 **99.17**	19.25 **99.25**	19.30 **99.30**	19.33 **99.33**	19.36 **99.34**	19.37 **99.36**	19.38 **99.38**	19.39 **99.40**	19.40 **99.41**	19.41 **99.42**	19.42 **99.43**	19.43 **99.44**	19.44 **99.45**	19.45 **99.46**	19.46 **99.47**	19.47 **99.48**	19.47 **99.48**	19.48 **99.49**	19.49 **99.49**	19.49 **99.49**	19.50 **99.50**	19.50 **99.50**
3	10.13 **34.12**	9.55 **30.81**	9.28 **29.46**	9.12 **28.71**	9.01 **28.24**	8.94 **27.91**	8.88 **27.67**	8.84 **27.49**	8.81 **27.34**	8.78 **27.23**	8.76 **27.13**	8.74 **27.05**	8.71 **26.92**	8.69 **26.83**	8.66 **26.69**	8.64 **26.60**	8.62 **26.50**	8.60 **26.41**	8.58 **26.30**	8.57 **26.27**	8.56 **26.23**	8.54 **26.18**	8.54 **26.14**	8.53 **26.12**
4	7.71 **21.20**	6.94 **18.00**	6.59 **16.69**	6.39 **15.98**	6.26 **15.52**	6.16 **15.21**	6.09 **14.98**	6.04 **14.80**	6.00 **14.66**	5.96 **14.54**	5.93 **14.45**	5.91 **14.37**	5.87 **14.24**	5.84 **14.15**	5.80 **14.02**	5.77 **13.93**	5.74 **13.83**	5.71 **13.74**	5.70 **13.69**	5.68 **13.61**	5.66 **13.57**	5.65 **13.52**	5.64 **13.48**	5.63 **13.46**
5	6.61 **16.26**	5.79 **13.27**	5.41 **12.06**	5.19 **11.39**	5.05 **10.97**	4.95 **10.67**	4.88 **10.45**	4.82 **10.27**	4.78 **10.15**	4.74 **10.05**	4.70 **9.96**	4.68 **9.89**	4.64 **9.77**	4.60 **9.68**	4.56 **9.55**	4.53 **9.47**	4.50 **9.38**	4.46 **9.29**	4.44 **9.24**	4.42 **9.17**	4.40 **9.13**	4.38 **9.07**	4.37 **9.04**	4.36 **9.02**
6	5.99 **13.74**	5.14 **10.92**	4.76 **9.78**	4.53 **9.15**	4.39 **8.75**	4.28 **8.47**	4.21 **8.26**	4.15 **8.10**	4.10 **7.98**	4.06 **7.87**	4.03 **7.79**	4.00 **7.72**	3.96 **7.60**	3.92 **7.52**	3.87 **7.39**	3.84 **7.31**	3.81 **7.23**	3.77 **7.14**	3.75 **7.09**	3.72 **7.02**	3.71 **6.99**	3.69 **6.94**	3.68 **6.90**	3.67 **6.88**
7	5.59 **12.25**	4.74 **9.55**	4.35 **8.45**	4.12 **7.85**	3.97 **7.46**	3.87 **7.19**	3.79 **7.00**	3.73 **6.84**	3.68 **6.71**	3.63 **6.62**	3.60 **6.54**	3.57 **6.47**	3.52 **6.35**	3.49 **6.27**	3.44 **6.15**	3.41 **6.07**	3.38 **5.98**	3.34 **5.90**	3.32 **5.85**	3.29 **5.78**	3.28 **5.75**	3.25 **5.70**	3.24 **5.67**	3.23 **5.65**
8	5.32 **11.26**	4.46 **8.65**	4.07 **7.59**	3.84 **7.01**	3.69 **6.63**	3.58 **6.37**	3.50 **6.19**	3.44 **6.03**	3.39 **5.91**	3.34 **5.82**	3.31 **5.74**	3.28 **5.67**	3.23 **5.56**	3.20 **5.48**	3.15 **5.36**	3.12 **5.28**	3.08 **5.20**	3.05 **5.11**	3.03 **5.06**	3.00 **5.00**	2.98 **4.96**	2.96 **4.91**	2.94 **4.88**	2.93 **4.86**
9	5.12 **10.56**	4.26 **8.02**	3.86 **6.99**	3.63 **6.42**	3.48 **6.06**	3.37 **5.80**	3.29 **5.62**	3.23 **5.47**	3.18 **5.35**	3.13 **5.26**	3.10 **5.18**	3.07 **5.11**	3.02 **5.00**	2.98 **4.92**	2.93 **4.80**	2.90 **4.73**	2.86 **4.64**	2.82 **4.56**	2.80 **4.51**	2.77 **4.45**	2.76 **4.41**	2.73 **4.36**	2.72 **4.33**	2.71 **4.31**

*Reprinted, by permission, from Snedecor, *Statistical Methods*, Iowa State College Press, Iowa State College, Ames.

10	4.96 **10.04**	4.10 **7.56**	3.71 **6.55**	3.48 **5.99**	3.33 **5.64**	3.22 **5.39**	3.14 **5.21**	3.07 **5.06**	3.02 **4.95**	2.97 **4.85**	2.94 **4.78**	2.91 **4.71**	2.86 **4.60**	2.82 **4.52**	2.77 **4.41**	2.74 **4.33**	2.70 **4.25**	2.67 **4.17**	2.64 **4.12**	2.61 **4.05**	2.59 **4.01**	2.56 **3.96**	2.55 **3.93**	2.54 **3.91**
11	4.84 **9.65**	3.98 **7.20**	3.59 **6.22**	3.36 **5.67**	3.20 **5.32**	3.09 **5.07**	3.01 **4.88**	2.95 **4.74**	2.90 **4.63**	2.86 **4.54**	2.82 **4.46**	2.79 **4.40**	2.74 **4.29**	2.70 **4.21**	2.65 **4.10**	2.61 **4.02**	2.57 **3.94**	2.53 **3.86**	2.50 **3.80**	2.47 **3.74**	2.45 **3.70**	2.42 **3.66**	2.41 **3.62**	2.40 **3.60**
12	4.75 **9.33**	3.88 **6.93**	3.49 **5.95**	3.26 **5.41**	3.11 **5.06**	3.00 **4.82**	2.92 **4.65**	2.85 **4.50**	2.80 **4.39**	2.76 **4.30**	2.72 **4.22**	2.69 **4.16**	2.64 **4.05**	2.60 **3.98**	2.54 **3.86**	2.50 **3.78**	2.46 **3.70**	2.42 **3.61**	2.40 **3.56**	2.36 **3.49**	2.35 **3.46**	2.32 **3.41**	2.31 **3.38**	2.30 **3.36**
13	4.67 **9.07**	3.80 **6.70**	3.41 **5.74**	3.18 **5.20**	3.02 **4.86**	2.92 **4.62**	2.84 **4.44**	2.77 **4.30**	2.72 **4.19**	2.67 **4.10**	2.63 **4.02**	2.60 **3.96**	2.55 **3.85**	2.51 **3.78**	2.46 **3.67**	2.42 **3.59**	2.38 **3.51**	2.34 **3.42**	2.32 **3.37**	2.28 **3.30**	2.26 **3.27**	2.24 **3.21**	2.22 **3.18**	2.21 **3.16**
14	4.60 **8.86**	3.74 **6.51**	3.34 **5.56**	3.11 **5.03**	2.96 **4.69**	2.85 **4.46**	2.77 **4.28**	2.70 **4.14**	2.65 **4.03**	2.60 **3.94**	2.56 **3.86**	2.53 **3.80**	2.48 **3.70**	2.44 **3.62**	2.39 **3.51**	2.35 **3.43**	2.31 **3.34**	2.27 **3.26**	2.24 **3.21**	2.21 **3.14**	2.19 **3.11**	2.16 **3.06**	2.14 **3.02**	2.13 **3.00**
15	4.54 **8.68**	3.68 **6.36**	3.29 **5.42**	3.06 **4.89**	2.90 **4.56**	2.79 **4.32**	2.70 **4.14**	2.64 **4.00**	2.59 **3.89**	2.55 **3.80**	2.51 **3.73**	2.48 **3.67**	2.43 **3.56**	2.39 **3.48**	2.33 **3.36**	2.29 **3.29**	2.25 **3.20**	2.21 **3.12**	2.18 **3.07**	2.15 **3.00**	2.12 **2.97**	2.10 **2.92**	2.08 **2.89**	2.07 **2.87**
16	4.49 **8.53**	3.63 **6.23**	3.24 **5.29**	3.01 **4.77**	2.85 **4.44**	2.74 **4.20**	2.66 **4.03**	2.59 **3.89**	2.54 **3.78**	2.49 **3.69**	2.45 **3.61**	2.42 **3.55**	2.37 **3.45**	2.33 **3.37**	2.28 **3.25**	2.24 **3.18**	2.20 **3.10**	2.16 **3.01**	2.13 **2.96**	2.09 **2.89**	2.07 **2.86**	2.04 **2.80**	2.02 **2.77**	2.01 **2.75**
17	4.45 **8.40**	3.59 **6.11**	3.20 **5.18**	2.96 **4.67**	2.81 **4.34**	2.70 **4.10**	2.62 **3.93**	2.55 **3.79**	2.50 **3.68**	2.45 **3.59**	2.41 **3.52**	2.38 **3.45**	2.33 **3.35**	2.29 **3.27**	2.23 **3.16**	2.19 **3.08**	2.15 **3.00**	2.11 **2.92**	2.08 **2.86**	2.04 **2.79**	2.02 **2.76**	1.99 **2.70**	1.97 **2.67**	1.96 **2.65**
18	4.41 **8.28**	3.55 **6.01**	3.16 **5.09**	2.93 **4.58**	2.77 **4.25**	2.66 **4.01**	2.58 **3.85**	2.51 **3.71**	2.46 **3.60**	2.41 **3.51**	2.37 **3.44**	2.34 **3.37**	2.29 **3.27**	2.25 **3.19**	2.19 **3.07**	2.15 **3.00**	2.11 **2.91**	2.07 **2.83**	2.04 **2.78**	2.00 **2.71**	1.98 **2.68**	1.95 **2.62**	1.93 **2.59**	1.92 **2.57**
19	4.38 **8.18**	3.52 **5.93**	3.13 **5.01**	2.90 **4.50**	2.74 **4.17**	2.63 **3.94**	2.55 **3.77**	2.48 **3.63**	2.43 **3.52**	2.38 **3.43**	2.34 **3.36**	2.31 **3.30**	2.26 **3.19**	2.21 **3.12**	2.15 **3.00**	2.11 **2.92**	2.07 **2.84**	2.02 **2.76**	2.00 **2.70**	1.96 **2.63**	1.94 **2.60**	1.91 **2.54**	1.90 **2.51**	1.88 **2.49**
20	4.35 **8.10**	3.49 **5.85**	3.10 **4.94**	2.87 **4.43**	2.71 **4.10**	2.60 **3.87**	2.52 **3.71**	2.45 **3.56**	2.40 **3.45**	2.35 **3.37**	2.31 **3.30**	2.28 **3.23**	2.23 **3.13**	2.18 **3.05**	2.12 **2.94**	2.08 **2.86**	2.04 **2.77**	1.99 **2.69**	1.96 **2.63**	1.92 **2.56**	1.90 **2.53**	1.87 **2.47**	1.85 **2.44**	1.84 **2.42**
21	4.32 **8.02**	3.47 **5.78**	3.07 **4.87**	2.84 **4.37**	2.68 **4.04**	2.57 **3.81**	2.49 **3.65**	2.42 **3.51**	2.37 **3.40**	2.32 **3.31**	2.28 **3.24**	2.25 **3.17**	2.20 **3.07**	2.15 **2.99**	2.09 **2.88**	2.05 **2.80**	2.00 **2.72**	1.96 **2.63**	1.93 **2.58**	1.89 **2.51**	1.87 **2.47**	1.84 **2.42**	1.82 **2.38**	1.81 **2.36**
22	4.30 **7.94**	3.44 **5.72**	3.05 **4.82**	2.82 **4.31**	2.66 **3.99**	2.55 **3.76**	2.47 **3.59**	2.40 **3.45**	2.35 **3.35**	2.30 **3.26**	2.26 **3.18**	2.23 **3.12**	2.18 **3.02**	2.13 **2.94**	2.07 **2.83**	2.03 **2.75**	1.98 **2.67**	1.93 **2.58**	1.91 **2.53**	1.87 **2.46**	1.84 **2.42**	1.81 **2.37**	1.80 **2.33**	1.78 **2.31**
23	4.28 **7.88**	3.42 **5.66**	3.03 **4.76**	2.80 **4.26**	2.64 **3.94**	2.53 **3.71**	2.45 **3.54**	2.38 **3.41**	2.32 **3.30**	2.28 **3.21**	2.24 **3.14**	2 20 **3.07**	2.14 **2.97**	2.10 **2.89**	2.04 **2.78**	2.00 **2.70**	1.96 **2.62**	1.91 **2.53**	1.88 **2.48**	1.84 **2.41**	1.82 **2.37**	1.79 **2.32**	1.77 **2.28**	1.76 **2.26**
24	4.26 **7.82**	3.40 **5.61**	3.01 **4.72**	2.78 **4.22**	2.62 **3.90**	2.51 **3.67**	2.43 **3.50**	2.36 **3.36**	2.30 **3.25**	2.26 **3.17**	2.22 **3.09**	2.18 **3.03**	2.13 **2.93**	2.09 **2.85**	2.02 **2.74**	1.98 **2.66**	1.94 **2.58**	1.89 **2.49**	1.86 **2.44**	1.82 **2.36**	1.80 **2.33**	1.76 **2.27**	1.74 **2.23**	1.73 **2.21**
25	4.24 **7.77**	3.38 **5.57**	2.99 **4.68**	2.76 **4.18**	2.60 **3.86**	2.49 **3.63**	2.41 **3.46**	2.34 **3.32**	2.28 **3.21**	2.24 **3.13**	2.20 **3.05**	2.16 **2.99**	2.11 **2.89**	2.06 **2.81**	2.00 **2.70**	1.96 **2.62**	1.92 **2.54**	1.87 **2.45**	1.84 **2.40**	1.80 **2.32**	1.77 **2.29**	1.74 **2.23**	1.72 **2.19**	1.71 **2.17**
26	4.22 **7.72**	3.37 **5.53**	2.89 **4.64**	2.74 **4.14**	2.59 **3.82**	2.47 **3.59**	2.39 **3.42**	2.32 **3.29**	2.27 **3.17**	2.22 **3.09**	2.18 **3.02**	2.15 **2.96**	2.10 **2.86**	2.05 **2.77**	1.99 **2.66**	1.95 **2.58**	1.90 **2.50**	1.85 **2.41**	1.82 **2.36**	1.78 **2.28**	1.76 **2.25**	1.72 **2.19**	1.70 **2.15**	1.69 **2.13**

F Distribution (*continued*)

5% (Roman Type) and 1% (Bold-Face Type) Points for the Distribution of *F*

Degrees of freedom for lesser mean square	Degrees of freedom for greater mean square																							
	1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	40	50	75	100	200	500	∞
27	4.21 **7.68**	3.35 **5.49**	2.96 **4.60**	2.73 **4.11**	2.57 **3.79**	2.46 **3.56**	2.37 **3.39**	2.30 **3.26**	2.25 **3.14**	2.20 **3.06**	2.16 **2.98**	2.13 **2.93**	2.08 **2.83**	2.03 **2.74**	1.97 **2.63**	1.93 **2.55**	1.88 **2.47**	1.84 **2.38**	1.80 **2.33**	1.76 **2.25**	1.74 **2.21**	1.71 **2.16**	1.68 **2.12**	1.67 **2.10**
28	4.20 **7.64**	3.34 **5.45**	2.95 **4.57**	2.71 **4.07**	2.56 **3.76**	2.44 **3.53**	2.36 **3.36**	2.29 **3.23**	3.24 **3.11**	2.19 **3.03**	2.15 **2.95**	2.12 **2.90**	2.06 **2.80**	2.02 **2.71**	1.96 **2.60**	1.91 **2.52**	1.87 **2.44**	1.81 **2.35**	1.78 **2.30**	1.75 **2.22**	1.72 **2.18**	1.69 **2.13**	1.67 **2.09**	1.65 **2.06**
29	4.18 **7.60**	3.33 **5.52**	2.93 **4.54**	2.70 **4.04**	2.54 **3.73**	2.43 **3.50**	2.35 **3.33**	2.28 **3.20**	2.22 **3.08**	2.18 **3.00**	2.14 **2.92**	2.10 **2.87**	2.05 **2.77**	2.00 **2.68**	1.94 **2.57**	1.90 **2.49**	1.85 **2.41**	1.80 **2.32**	1.77 **2.27**	1.73 **2.19**	1.71 **2.15**	1.68 **2.10**	1.65 **2.06**	1.64 **2.03**
30	4.17 **7.56**	3.32 **5.39**	2.92 **4.51**	2.69 **4.02**	2.53 **3.70**	2.42 **3.47**	2.34 **3.30**	2.27 **3.17**	2.21 **3.06**	2.16 **2.98**	2.12 **2.90**	2.09 **2.84**	2.04 **2.74**	1.99 **2.66**	1.93 **2.55**	1.89 **2.47**	1.84 **2.38**	1.79 **2.29**	1.76 **2.24**	1.72 **2.16**	1.69 **2.13**	1.66 **2.07**	1.64 **2.03**	1.62 **2.01**
32	4.15 **7.50**	3.30 **5.34**	2.90 **4.46**	2.67 **3.97**	2.51 **3.66**	2.40 **3.42**	2.32 **3.25**	2.25 **3.12**	2.19 **3.01**	2.14 **2.94**	2.10 **2.86**	2.07 **2.80**	2.02 **2.70**	1.97 **2.62**	1.91 **2.51**	1.86 **2.42**	1.82 **2.34**	1.76 **2.25**	1.74 **2.20**	1.69 **2.12**	1.67 **2.08**	1.64 **2.02**	1.61 **1.98**	1.59 **1.96**
34	4.13 **7.44**	3.28 **5.29**	2.88 **4.42**	2.65 **3.93**	2.49 **3.61**	2.38 **3.38**	2.30 **3.21**	2.23 **3.08**	2.17 **2.97**	2.12 **2.89**	2.08 **2.82**	2.05 **2.76**	2.00 **2.66**	1.95 **2.58**	1.89 **2.47**	1.84 **2.38**	1.80 **2.30**	1.74 **2.21**	1.71 **2.15**	1.67 **2.08**	1.64 **2.01**	1.61 **1.98**	1.59 **1.94**	1.57 **1.91**
36	4.11 **7.39**	3.26 **5.25**	2.86 **4.38**	2.63 **3.89**	2.48 **3.58**	2.36 **3.35**	2.28 **3.18**	2.21 **3.04**	2.15 **2.94**	2.10 **2.86**	2.06 **2.78**	2.03 **2.72**	1.89 **2.62**	1.93 **2.54**	1.87 **2.43**	1.82 **2.35**	1.78 **2.26**	1.72 **2.17**	1.69 **2.12**	1.65 **2.04**	1.62 **2.00**	1.59 **1.94**	1.56 **1.90**	1.55 **1.87**
38	4.10 **7.35**	3.25 **5.21**	2.85 **4.34**	2.62 **3.86**	2.46 **3.54**	2.35 **3.32**	2.26 **3.15**	2.19 **3.02**	2.14 **2.91**	2.09 **2.82**	2.05 **2.75**	2.02 **2.69**	1.96 **2.59**	1.92 **2.51**	1.85 **2.40**	1.80 **2.32**	1.76 **2.22**	1.71 **2.14**	1.67 **2.08**	1.63 **2.00**	1.60 **1.97**	1.57 **1.90**	1.54 **1.86**	1.53 **1.84**
40	4.08 **7.31**	3.23 **5.18**	2.84 **4.31**	2.61 **3.83**	2.45 **3.51**	2.34 **3.29**	2.25 **3.12**	2.18 **2.99**	2.12 **2.88**	2.07 **2.80**	2.04 **2.73**	2.00 **2.66**	1.95 **2.56**	1.90 **2.49**	1.84 **2.37**	1.79 **2.29**	1.74 **2.20**	1.69 **2.11**	1.66 **2.05**	1.61 **1.97**	1.59 **1.94**	1.55 **1.88**	1.53 **1.84**	1.51 **1.81**
42	4.07 **7.27**	3.22 **5.15**	2.83 **4.29**	2.59 **3.80**	2.44 **3.49**	2.32 **3.26**	2.24 **3.10**	2.17 **2.96**	2.11 **2.86**	2.06 **2.77**	2.02 **2.70**	1.99 **2.64**	1.94 **2.54**	1.89 **2.46**	1.82 **2.35**	1.78 **2.26**	1.73 **2.17**	1.68 **2.08**	1.64 **2.02**	1.60 **1.94**	1.57 **1.91**	1.54 **1.85**	1.51 **1.80**	1.49 **1.78**
44	4.06 **7.24**	3.21 **5.12**	2.82 **4.26**	2.58 **3.78**	2.43 **3.46**	2.31 **3.24**	2.23 **3.07**	2.16 **2.94**	2.10 **2.84**	2.05 **2.75**	2.01 **2.68**	1.98 **2.62**	1.92 **2.52**	1.88 **2.44**	1.81 **2.32**	1.76 **2.24**	1.72 **2.15**	1.66 **2.06**	1.63 **2.00**	1.58 **1.92**	1.56 **1.88**	1.52 **1.82**	1.50 **1.78**	1.48 **1.75**
46	4.05 **7.21**	3.20 **5.10**	2.81 **4.24**	2.57 **3.76**	2.42 **3.44**	2.30 **3.22**	2.22 **3.05**	2.14 **2.92**	2.09 **2.82**	2.04 **2.73**	2.00 **2.66**	1.97 **2.60**	1.91 **2.50**	1.87 **2.42**	1.80 **2.30**	1.75 **2.22**	1.71 **2.13**	1.65 **2.04**	1.62 **1.98**	1.57 **1.90**	1.54 **1.86**	1.51 **1.80**	1.48 **1.76**	1.46 **1.72**
48	4.04 **7.19**	3.19 **5.08**	2.80 **4.22**	2.56 **3.74**	2.41 **3.42**	2.30 **3.20**	2.21 **3.04**	2.14 **2.90**	2.08 **2.80**	2.03 **2.71**	1.99 **2.64**	1.96 **2.58**	1.90 **2.48**	1.86 **2.40**	1.79 **2.28**	1.74 **2.20**	1.70 **2.11**	1.64 **2.02**	1.61 **1.96**	1.56 **1.88**	1.53 **1.84**	1.50 **1.78**	1.47 **1.73**	1.45 **1.70**

50	4.03 **7.17**	3.18 **5.06**	2.79 **4.20**	2.56 **3.72**	2.40 **3.41**	2.29 **3.18**	2.20 **3.02**	2.13 **2.88**	2.07 **2.78**	2.02 **2.70**	1.98 **2.62**	1.95 **2.56**	1.90 **2.46**	1.85 **2.39**	1.78 **2.26**	1.74 **2.18**	1.69 **2.10**	1.63 **2.00**	1.60 **1.94**	1.55 **1.86**	1.52 **1.82**	1.48 **1.76**	1.46 **1.71**	1.44 **1.68**
55	4.02 **7.12**	3.17 **5.01**	2.78 **4.16**	2.54 **3.68**	2.38 **3.37**	2.27 **3.15**	2.18 **2.98**	2.11 **2.85**	2.05 **2.75**	2.00 **2.66**	1.97 **2.59**	1.93 **2.53**	1.88 **2.43**	1.83 **2.35**	1.76 **2.23**	1.72 **2.15**	1.67 **2.06**	1.61 **1.96**	1.58 **1.90**	1.52 **1.82**	1.50 **1.78**	1.46 **1.71**	1.43 **1.66**	1.41 **1.64**
60	4.00 **7.08**	3.15 **4.98**	2.76 **4.13**	2.52 **3.65**	2.37 **3.34**	2.25 **3.12**	2.17 **2.95**	2.10 **2.82**	2.04 **2.72**	1.99 **2.63**	1.95 **2.56**	1.92 **2.50**	1.86 **2.40**	1.81 **2.32**	1.75 **2.20**	1.70 **2.12**	1.65 **2.03**	1.59 **1.93**	1.56 **1.87**	1.50 **1.79**	1.48 **1.74**	1.44 **1.68**	1.41 **1.63**	1.39 **1.60**
65	3.99 **7.04**	3.14 **4.95**	2.75 **4.10**	2.51 **3.62**	2.36 **3.31**	2.24 **3.09**	2.15 **2.93**	2.08 **2.79**	2.02 **2.70**	1.98 **2.61**	1.94 **2.54**	1.90 **2.47**	1.85 **2.37**	1.80 **2.30**	1.73 **2.18**	1.68 **2.09**	1.63 **2.00**	1.57 **1.90**	1.54 **1.84**	1.49 **1.76**	1.46 **1.71**	1.42 **1.64**	1.39 **1.60**	1.37 **1.56**
70	3.98 **7.01**	3.13 **4.92**	2.74 **4.08**	2.50 **3.60**	2.35 **3.29**	2.32 **3.07**	2.14 **2.91**	2.07 **2.77**	2.01 **2.67**	1.97 **2.59**	1.93 **2.51**	1.89 **2.45**	1.84 **2.35**	1.79 **2.28**	1.72 **2.15**	1.67 **2.07**	1.62 **1.98**	1.56 **1.88**	1.53 **1.82**	1.47 **1.74**	1.45 **1.69**	1.40 **1.62**	1.37 **1.56**	1.35 **1.53**
80	3.96 **6.96**	3.11 **4.88**	2.72 **4.04**	2.48 **3.56**	2.33 **3.25**	2.21 **3.04**	2.12 **2.87**	2.05 **2.74**	1.99 **2.64**	1.95 **2.55**	1.91 **2.48**	1.88 **2.41**	1.82 **2.32**	1.77 **2.24**	1.70 **2.11**	1.65 **2.03**	1.60 **1.94**	1.54 **1.84**	1.51 **1.78**	1.45 **1.70**	1.42 **1.65**	1.38 **1.57**	1.35 **1.52**	1.32 **1.49**
100	3.94 **6.90**	3.09 **4.82**	2.70 **3.98**	2.46 **3.51**	2.30 **3.20**	2.19 **2.99**	2.10 **2.82**	2.03 **2.69**	1.97 **2.59**	1.92 **2.51**	1.88 **2.43**	1.85 **2.36**	1.79 **2.26**	1.75 **2.19**	1.68 **2.06**	1.63 **1.98**	1.57 **1.89**	1.51 **1.79**	1.48 **1.73**	1.42 **1.64**	1.39 **1.59**	1.34 **1.51**	1.30 **1.46**	1.28 **1.43**
125	3.92 **6.84**	3.07 **4.78**	2.68 **3.94**	2.44 **3.47**	2.29 **3.17**	2.17 **2.95**	2.08 **2.79**	2.01 **2.65**	1.95 **2.56**	1.90 **2.47**	1.86 **2.40**	1.83 **2.33**	1.77 **2.23**	1.72 **2.15**	1.65 **2.03**	1.60 **1.94**	1.55 **1.85**	1.49 **1.75**	1.45 **1.68**	1.39 **1.59**	1.36 **1.54**	1.31 **1.46**	1.27 **1.40**	1.25 **1.37**
150	3.91 **6.81**	3.06 **4.75**	2.67 **3.91**	2.43 **3.44**	2.27 **3.13**	2.16 **2.92**	2.07 **2.76**	2.00 **2.62**	1.94 **2.53**	1.89 **2.44**	1.85 **2.37**	1.82 **2.30**	1.76 **2.20**	1.71 **2.12**	1.64 **2.00**	1.59 **1.91**	1.54 **1.83**	1.47 **1.72**	1.44 **1.66**	1.37 **1.56**	1.34 **1.51**	1.29 **1.43**	1.25 **1.37**	1.22 **1.33**
200	3.89 **6.76**	3.04 **4.71**	2.65 **3.88**	2.41 **3.41**	2.26 **3.11**	2.14 **2.90**	2.05 **2.73**	1.98 **2.60**	1.92 **2.50**	1.87 **2.41**	1.83 **2.34**	1.80 **2.28**	1.74 **1.17**	1.69 **2.09**	1.62 **1.97**	1.57 **1.88**	1.52 **1.79**	1.45 **1.69**	1.42 **1.62**	1.35 **1.53**	1.32 **1.48**	1.26 **1.39**	1.22 **1.33**	1.19 **1.28**
400	3.86 **6.70**	3.02 **4.66**	2.62 **3.83**	2.39 **3.36**	2.23 **3.06**	2.12 **2.85**	2.03 **2.69**	1.96 **2.55**	1.90 **2.46**	1.85 **2.37**	1.81 **2.29**	1.78 **2.23**	1.72 **2.12**	1.67 **2.04**	1.60 **1.92**	1.54 **1.84**	1.49 **1.74**	1.42 **1.64**	1.38 **1.57**	1.32 **1.47**	1.28 **1.42**	1.22 **1.32**	1.16 **1.24**	1.13 **1.19**
1000	3.85 **6.66**	3.00 **4.62**	2.61 **3.80**	2.38 **3.34**	2.22 **3.04**	2.10 **2.82**	2.02 **2.66**	1.95 **2.53**	1.89 **2.43**	1.84 **2.34**	1.80 **2.26**	1.76 **2.20**	1.70 **2.09**	1.65 **2.01**	1.58 **1.89**	1.53 **1.81**	1.47 **1.71**	1.41 **1.61**	1.36 **1.54**	1.30 **1.44**	1.26 **1.38**	1.19 **1.28**	1.13 **1.19**	1.08 **1.11**
∞	3.84 **6.64**	2.99 **4.60**	2.60 **3.78**	2.37 **3.32**	2.21 **3.02**	2.09 **2.80**	2.01 **2.64**	1.94 **2.51**	1.88 **2.41**	1.83 **2.32**	1.79 **2.24**	1.75 **2.18**	1.69 **2.07**	1.64 **1.99**	1.57 **1.87**	1.52 **1.79**	1.46 **1.69**	1.40 **1.59**	1.35 **1.52**	1.28 **1.41**	1.24 **1.36**	1.17 **1.25**	1.11 **1.15**	1.00 **1.00**

Index